Genetic variation in DNA repair proteins modifies the course of Huntington's disease

Thesis submitted for the degree of Doctor of Philosophy

Michael Flower

Institute of Neurology

University College London

2018

For Chloe and my parents

Declaration

I, Michael Flower, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

Above all, I would like to thank the patients and their families; without their involvement and dedication to research this work would not have been possible. I hope these findings contribute to the search for curative treatments for repeat expansion diseases such as Huntington's.

I would like to thank my supervisors Sarah Tabrizi, Gillian Bates and Lesley Jones, who have provided me with exceptional support, mentorship and guidance. I have learnt more from them than I could possibly express.

I would also like to thank Rob Goold, Davina Hensman-Moss and Ralph Andre and who have provided constant support and mentorship. I extend this thanks to the lab team at the UCL Huntington's Disease Centre, particularly Jamie Miller, Rhia Ghosh and Alison Wood-Kaczmar.

Finally, I would like to thank my fiancé Chloe, the driving force behind everything I've achieved, my late mother for being an endless inspiration and my father, for whom nothing is too much to ask.

Abstract

Huntington's disease (HD) is caused by a CAG repeat expansion in *HTT* on chromosome 4. Onset and progression are inversely correlated with repeat length, but a significant proportion of the variability in each is due to modifiers elsewhere in the genome. Recent genome-wide association studies have identified the DNA repair genes *FAN1* and *MSH3* as modifiers of onset and progression respectively. This thesis finds that variants associated with HD disease course also influence onset in other polyglutamine diseases, suggesting a shared pathogenic mechanism involving DNA repair. In blood from HD patients there is significant transcriptional dysregulation, particularly involving immune, metabolic and DNA repair pathways, which correlates with disease severity, parallels dysregulation seen in the most affected HD brain regions and overlaps with Alzheimer's disease. To study the role of DNA repair, several cell models of somatic instability were developed, including patient-derived lymphoblasts and induced pluripotent stem cells, which show exponential repeat expansion that continues in differentiated medium spiny neurons (MSNs). In a *FAN1* knockout U20S cell model of HD, *FAN1* is shown to protect against repeat instability, and this function is dependent on protein concentration and CAG repeat length, but does not require its nuclease activity. shRNA-mediated *FAN1* knockdown accelerates repeat expansion in both patient-derived iPSCs and MSNs. Through chromatin immunoprecipitation, FAN1 is shown to bind, but not specifically target, CAG repeat DNA. AAV9-mediated miRNA *Fan1* knockdown in the striatum and liver of R6/2 mice did not accelerate repeat expansion, likely because only 23% knockdown was achieved. Illumina sequencing of the *MSH3* region that influences HD progression identified a repeat variant that is associated with decreased *MSH3* expression, reduced somatic expansion, delayed onset and slower progression in HD and myotonic dystrophy type 1 (DM1). These results suggest *MSH3* promotes and *FAN1* protects against repeat instability, which in turn influences the course of repeat expansion diseases.

Impact statement

This thesis suggests that in repeat expansion diseases, a network of DNA repair proteins causes somatic instability, which in turn influences disease course. Modulation of DNA repair components, such as FAN1 and MSH3, has significant therapeutic potential in some of the commonest genetic neurodegenerative disorders through the reduction of somatic expansion. The blood transcriptomic signature of HD, with dysregulation of the immune system, DNA repair, RNA processing and energy metabolism also overlaps with that of Alzheimer's disease. Several synthetic and patient-derived Huntington's disease cell models of repeat length-dependent somatic instability were generated, including medium spiny neurons, the cells most vulnerable to the disease. These will permit the elucidation of the wider DNA repair network responsible and the investigation of compounds that reduce instability. FAN1 has been shown for the first time to bind CAG repeat DNA and increasing its expression may protect against expansion through a novel mechanism that is independent of its nuclease activity. Conversely, decreasing expression of *MSH3* may reduce somatic expansion, delay onset and slow progression in several repeat expansion diseases.

Table of Contents

Declaration ***3***

Acknowledgements ***4***

Abstract ***5***

Impact statement ***5***

Figures ***10***

Tables ***15***

Abbreviations ***18***

Chapter 1 Introduction ***26***

1.1 DNA repeat expansion 26

1.2 Huntington's disease 28

1.3 DNA repair 37

1.4 FAN1 42

1.5 MSH3 46

1.6 The immune system in neurodegenerative disease 51

Chapter 2 Materials and methods ***53***

2.1 Cell lines 53

2.2 Cell culture 62

2.3 Cell imaging 68

2.4 Genetics 68

2.5 Protein 82

Chapter 3 DNA repair variants modify phenotype in polyglutamine diseases ***83***

3.1 Background 83

3.2 Aims 87

3.3 Methods 88

3.4 Contributions 91

3.5 Results 92

3.6 Discussion 95

3.7 Summary 96
3.8 Publications relating to this chapter 96
Chapter 4 Transcriptional dysregulation in Huntington's disease patient blood 97
4.1 Background 97
4.2 Aim 100
4.3 Methods 100
4.4 Contributions 103
4.5 Results 105
4.6 Discussion 121
4.7 Summary 124
4.8 Publications relating to this chapter 124
Chapter 5 Cell models of HTT CAG repeat instability 125
5.1 Background 125
5.2 Aims 128
5.3 Methods 130
5.4 Contributions 132
5.5 Results 133
5.6 Discussion 158
5.7 Summary 160
5.8 Publications relating to this chapter 160
Chapter 6 FAN1 activity at HTT CAG repeat DNA 161
6.1 Background 161
6.2 Aims 162
6.3 Methods 163
6.4 Contributions 168
6.5 Results 169
6.6 Discussion 190
6.7 Summary 192
6.8 Publications relating to this chapter 193

Chapter 7 Fan1 knockdown in R6/2 mice 194

7.1 Background 194

7.2 Aims 196

7.3 Methods 196

7.4 Contributions 203

7.5 Results 204

7.6 Discussion 235

7.7 Summary 236

Chapter 8 MSH3 modifies somatic instability and disease severity in Huntington's disease and myotonic dystrophy type 1 238

8.1 Background 238

8.2 Aims 239

8.3 Methods 240

8.4 Contributions 244

8.5 Results 245

8.6 Discussion 264

8.7 Summary 265

8.8 Publications related to this chapter 266

Chapter 9 Conclusions and future work 267

9.1 Conclusions 267

9.2 Future work 273

Chapter 10 Appendix 276

10.1 p'HRsincpptUCOE+htt exon1 IRES eGFP 129CAG vector sequence 276

10.2 U20S curve modelling in R 278

10.3 pSUPER.retro.puro vector sequence 279

10.4 pcDNA5-GFP-FAN1/FRT/TO vector sequence 280

10.5 Sequencing of A2UCOE *HTT* exon 1 construct CAG repeat regions 282

10.6 *MSH3* MiSeq primer sequences 283

10.7 PhIX reference sequence 284

10.8 MSH3 reference sequences 285
10.9 MiSeq library quality control Galaxy workflow 295
10.10 MSH3 repeat genotyping Galaxy workflow 302
10.11 MSH3 variant calling Galaxy workflow 305
10.12 Base-wise conservation scores across the MSH3 9bp tandem repeat region 307
10.13 MSH3 and DHFR expression quantitative trait loci (eQTL) 309
10.14 HD transcriptome-wide association study (TWAS) 313
Publications relating to this thesis 322
Funding 323
References 324

Figures

Figure 1.1. Potential future therapeutic targets in Huntington's disease. 33
Figure 1.2. Relationship between expanded CAG repeat length (x axis) and onset of diagnostic motor signs (y axis). 34
Figure 1.3. Manhattan plot of meta-analysis from GeM GWAS of HD motor onset. 36
Figure 1.4. Schematic representation of FAN1. 43
Figure 1.5. FAN1 interactome. 45
Figure 1.6. Schematic representation of MSH3. 47
Figure 1.7. MSH3 protein interactions. 48
Figure 1.8. Alignment of mouse and human MSH3 protein sequences. 50
Figure 1.9. Conservation of the MSH3 N-terminal domain between mouse and human. 51
Figure 2.1. p'HRsincpptUCOE+htt exon1 IRES eGFP 129CAG vector. 54
Figure 2.2. Micrographs of ReN VM cells transduced to express HTT exon 1 with 129 CAG repeats and GFP. 54
Figure 2.3. Schematic representation of FAN1. 57
Figure 2.4. CAG repeat sizing in 250Q lymphoblasts from Nance et al. (1999). 58
Figure 2.5. BD vacutainer centrifugation. 59
Figure 2.6. Micrographs of 125Q pluripotent ESCs. 60
Figure 2.7. Representative light micrograph of 109Q iPSCs (5x). 61
Figure 2.8. pcDNA5/FRT/TO vector containing Cas9-Flag under a tetracycline-inducible promoter and a hygromycin resistance gene. 61
Figure 2.9. Light micrograph of U20S FAN$^{-/-}$ cells in culture. 62
Figure 2.10. Neuronal differentiation of ReN VM cells expressing HTT exon 1 with 129 CAG repeats. 63
Figure 2.11. Neuronal differentiation of ReN CX cells expressing HTT exon 1 with 129 CAG repeats. 64
Figure 2.12. Scratch pattern 1 for MSN differentiation passage 1. 66
Figure 2.13. Scratch pattern 2 for MSN differentiation passage 2. 66
Figure 2.14. Light micrographs of differentiated medium spiny neurons (MSN). 67
Figure 2.15. Light micrograph of 109Q neural stem cells. 68
Figure 2.16. Primers for CAG repeat sizing. 70
Figure 2.17. Boxplot of variability in CAG sizing from cell lines with a range of repeat lengths. 72
Figure 2.18. Change in modal CAG repeat length. 73
Figure 2.19. Instability index calculation from Lee et al. (2010). 74
Figure 2.20. Proportional expansion analysis. 75
Figure 2.21. Representative example of fragment analysis traces from mouse #79 for the tissues and ages indicated (left). 75
Figure 2.22. Schematic representation of HTT primers on the genomic sequence. 79
Figure 2.23. HTT CAG repeat primers marked on the genomic sequence. 80
Figure 2.24. Schematic representation of ATXN3 primers on the genomic sequence. 80
Figure 2.25. Schematic representation of DMPK primers on the genomic sequence. 81

Figure 2.26. Schematic representation of FXN primers on the genomic sequence. 81
Figure 2.27. Schematic representation of TBP primers on the genomic sequence. 81
Figure 4.1. Upregulated pathways in HD versus control blood. 107
Figure 4.2. Downregulated pathways in HD versus control blood. 107
Figure 5.1. Cloning FAN1 shRNA into the pSUPER.retro.puro vector. 131
Figure 5.2. Oxidative stress in ReNeuron VM 129Q cells. 133
Figure 5.3. Representative CAG repeat sizing from ReN VM neural stem cells (NSC) differentiated for 56 days. 134
Figure 5.4. Repeat expansion analysis in ReN VM cells cultured as neural stem cells (NSC) or differentiated (MSN) for 56 days. 134
Figure 5.5. Representative CAG repeat sizing from ReN VM 129Q neural stem cells (NSC) chronically stressed with H_2O_2 during differentiation for 48 days. 135
Figure 5.6. Representative CAG repeat sizing from ReN VM 129Q NSCs 44d after initiation of differentiation, chronically stressed with H_2O_2 from day 15. 135
Figure 5.7. Representative CAG repeat sizing in ReN CX 129Q cells differentiated in the presence of chronic oxidative stress. 136
Figure 5.8. Representative CAG repeat sizing in ReN CX 129Q cells differentiated for 14 days. 137
Figure 5.9. CAG repeat sizing in ReN VM 129Q single cell clones (SCC). 138
Figure 5.10. CAG expansion analysis of ReN VM 129Q single cell clones. 139
Figure 5.11. Oxidative stress in HD lymphoblastoid (LB) cells. 140
Figure 5.12. Representative CAG repeat sizing in lymphoblastoid (LB) cells chronically stressed with the indicated H_2O_2 concentration. 141
Figure 5.13. Repeat expansion analysis in a 43Q, 44Q, 52Q and p.R507H LB lines chronically exposed to oxidative stress. 142
Figure 5.14. CAG repeat sizing in 250Q lymphoblasts (LB) by TP-PCR capillary electrophoresis. 143
Figure 5.15. Oxidative stress in QS3.2 and 109Q iPSCs. 143
Figure 5.16. shRNA mediated FAN1 knockdown in QS3.2 and 109Q iPSCs. 144
Figure 5.17. Baseline CAG repeat sizing in QS3 iPSCs. 145
Figure 5.18. CAG expansion analysis in QS3 iPSCs in culture, chronic oxidative stress and differentiation as NSCs or MSNs. 146
Figure 5.19. CAG repeat sizing of whole blood from a 125Q HD subject sampled 3 years apart. 147
Figure 5.20. CAG repeat sizing of lymphoblasts (LB) from a subject with 125 CAG repeats at baseline and following the emergence of a clone. 148
Figure 5.21. CAG repeat expansion analysis in 125 CAG LB cells. 149
Figure 5.22. CAG repeat sizing in 125Q iPSCs. 150
Figure 5.23. Exponential model of modal CAG repeat expansion in 109Q iPSCs. 151
Figure 5.24. CAG repeat expansion in 109Q iPSCs exposed to chronic oxidative stress. 152
Figure 5.25. Immunofluorescence confocal microscopy of differentiated 109Q medium spiny neurons (MSNs) treated with either FAN1 knockdown, empty vector or in control conditions. 153

Figure 5.26. Stable shRNA-mediated FAN1 knockdown in 109Q iPSCs and MSNs. 154
Figure 5.27. Comparison of exponential expansion models in 109Q iPSC, NSC and MSNs. 155
Figure 5.28. CAG repeat expansion in 109Q iPSCs and MSNs following shRNA-mediated FAN1 knockdown. 156
Figure 5.29. Comparison of CAG repeat expansion rate in HD cell lines. 157
Figure 6.1. Colony screening by restriction digest. 166
Figure 6.2. Transient transfection of HEK293 cells with full length Myc-tagged FAN1 variants. 169
Figure 6.3. Stable transfection of HEK293 cells. 170
Figure 6.4. Confocal immunofluorescence shows Myc-tagged FAN1 is expressed in the nucleus and forms repair foci that colocalise with FANCD2 following MMC. 171
Figure 6.5. siRNA mediated knockdown of endogenous FAN1 in HEK293 cells stably transfected with Myc-tagged p.R507H FAN1. 172
Figure 6.6. Patient-derived LB cell MMC sensitivity. 173
Figure 6.7. Immunoblot for FAN1 expression in HD lymphoblasts from subjects with the given rs3512 genotype. 174
Figure 6.8. Sanger sequencing confirming SDM of the pcDNA5/FRT/TO FAN1 vector. 175
Figure 6.9. Tetracycline induction reverses the protective effect of GFP-FAN1. 176
Figure 6.10. Tetracycline dose titration in p.R507H FAN1 U20S cells exposed to MMC. 177
Figure 6.11. γ-H2AX assay following cisplatin exposure in U20S cells. 178
Figure 6.12. FAN1 knockout sensitises U20S cells to MMC-induced interstrand crosslinks (ICL) and expression of wild type or variant GFP-FAN1 restores resistance. 179
Figure 6.13. Genotoxin assays in U20S cells. 180
Figure 6.14. Model of U20S FAN1$^{-/-}$ 118Q exponential expansion. 181
Figure 6.15. FAN1 protects against U20S 118Q CAG repeat expansion in a dose-dependent manner. 182
Figure 6.16. Model of U20S FAN1$^{-/-}$ 97Q exponential expansion. 183
Figure 6.17. CAG repeat expansion in U20S FAN1$^{-/-}$ cells is length dependent. 184
Figure 6.18. Change in modal CAG repeat length of U20S FAN1-/- cells expressing HTT exon 1 with 29-118Q. 184
Figure 6.19. Overlay of U20S FAN1$^{-/-}$ exponential expansion. 185
Figure 6.20. FAN1 nuclease and p.R507H variants do not modify CAG repeat expansion rate. 186
Figure 6.21. FAN1 binds HTT CAG repeat DNA. 187
Figure 6.22. FAN1 interacts with endogenous HTT DNA of 109Q iPSCs. 188
Figure 6.23. FAN1 interacts with endogenous HTT DNA of 125Q iPSCs. 188
Figure 6.24. FAN1 interacts with endogenous HTT DNA of HD lymphoblasts (LB). 189
Figure 6.25. Potential mechanisms by which FAN1 may protect against CAG repeat expansion. 192
Figure 7.1. miRNA design. 197
Figure 7.2. Fan1 silencing in 3T3 mouse embryonic fibroblasts (MEF) using AAV9 cB7 eGFP miRNA constructs. 198
Figure 7.3. Experimental protocol. 201
Figure 7.4. qPCR cycle threshold in pilot study of Fan1 expression in R6/2 tissues at 4 and 14 weeks. 205
Figure 7.5. Relative Fan1 expression level in pilot study of R6/2 tissues at 4 and 14 wk age. 206
Figure 7.6. Comparing DNA and RNA yield and purity from 3-in-1 and traditional extractions. 207

Figure 7.7. Comparing Fan1 expression level in 3-in-1 or traditional RNA-extracted cortex samples. 208
Figure 7.8. Western blot comparing Fan1 protein levels from 3-in-1 or traditional protein extractions from cortex. 209
Figure 7.9. A striatal sample (left) divided into thirds (right). 209
Figure 7.10. Fragment analysis from 1/3 of a striatum. Representative trace. 210
Figure 7.11. Protein extraction from 1/3 striatum of R6/2 mice. 210
Figure 7.12. RNA extraction from 1/3 striatum and Fan1 expression. 211
Figure 7.13. Fan1 knockdown in R6/2 following AAV9 cB7 eGFP.oligo 09 transduction. 212
Figure 7.14. Fan1 knockdown in R6/2 liver. 213
Figure 7.15. GFP expression following intrastriatal delivery of AAV9.mFan1 or AAV9.Scrambled control miRNA. 214
Figure 7.16. Representative sagittal sections showing GFP expression in the striatum. 214
Figure 7.17. GFP distribution pattern in R6/2 mice receiving intrastriatal injection of AAV9.mFan1.miRNA. 215
Figure 7.18. GFP expression in the liver. 215
Figure 7.19. GFP expression pattern across all study groups receiving IP injections 216
Figure 7.20. GFP intensity profile in transduced striatum and liver of R6/2 mice. 216
Figure 7.21. Mutant huntingtin aggregates in transduced R6/2 striatum. 217
Figure 7.22. Weight and temperature in 11-week mice. 218
Figure 7.23. Liver Fan1 expression. 219
Figure 7.24. Striatum Fan1 relative expression. 221
Figure 7.25. Striatum Fan1 expression. 223
Figure 7.26. Fan1 knockdown in R6/2 tissues. 224
Figure 7.27. Modal CAG repeat size at baseline (12 day tail). 225
Figure 7.28. Change in modal CAG repeat length relative to 12 day tail. 226
Figure 7.29. Somatic instability index relative to 12 day tail. 228
Figure 7.30. Proportional expansion analysis. 230
Figure 7.31. Change in modal CAG repeat length against Fan1 expression. 232
Figure 7.32. Somatic instability index against Fan1 expression. 233
Figure 7.33. Proportional expansion analysis against Fan1 expression. 234
Figure 8.1. Schematic of sequencing design for the MSH3 exon 1 region. 242
Figure 8.2. MSH3/DHFR 9bp tandem repeat allele structure and frequency observed in HD and DM1 cohorts. 246
Figure 8.3. Representative Sanger sequencing of a 3a heterozygote. 247
Figure 8.4. The MSH3 N-terminal region is poorly conserved between species. 248
Figure 8.5. The number of MSH3 3a repeat alleles is associated with HD and DM1 phenotypes. 249
Figure 8.6. Variants at the MSH3/DHFR locus are associated with phenotypes in HD and DM1. 253
Figure 8.7. MSH3 expression in 6a and 3a repeat homozygotes. 259
Figure 8.8. Association of the MSH3 3a allele with MSH3 and DHFR expression in HD whole blood. 260
Figure 8.9. MSH3 repeat length correlation with somatic expansion, age at onset, progression score and blood expression of MSH3 and DHFR in HD. 261

Figure 8.10. Association of the MSH3 3a allele with MSH3 and DHFR expression in the TRACK-HD prefrontal cortex TWAS. 262
Figure 8.11. MSH3 repeat length correlation with prefrontal cortex expression of MSH3 and DHFR in TRACK-HD. 263
Figure 8.12. CAG repeat expansion correlation with MSH3 or DHFR expression in TRACK-HD prefrontal cortex. 263
Figure 9.1. Potential mechanisms by which FAN1 may protect against CAG repeat expansion. 272
Figure 10.1. Schematic representation of pcDNA5-GFP-FAN1/FRT/TO vector 281
Figure 10.2. Sanger sequencing of A2UCOE HTT exon 1 construct CAG repeat regions. 282

Tables

Table 1.1 Genetic HD phenocopies. 31
Table 1.2. Acquired HD phenocopies. 31
Table 1.3. Notable studies proposing genetic modifiers of Huntington's disease. 35
Table 2.1. FAN1 variants identified by whole exome sequencing (WES) in fast and slow progressing subjects from TRACK-HD. 56
Table 2.2. Track-HD patient-derived lymphoblastoid (LB) cell lines used in this study. 56
Table 2.3. Calculation of CAG repeat size. 72
Table 2.4. Variability in CAG sizing from cell lines with a range of repeat lengths. 73
Table 2.5. HTT PCR primers. 79
Table 2.6. ATXN3 PCR primers. 80
Table 2.7. DMPK PCR primers. 80
Table 2.8. FXN PCR primers. 81
Table 2.9. TBP PCR primers. 81
Table 3.1. Characteristics of the polyglutamine diseases. 84
Table 3.2. Phenotypes of polyglutamine diseases. 85
Table 3.3. Cohort characteristics. 88
Table 3.4. Characteristics of single nucleotide polymorphisms (SNPs) used in this study. 89
Table 3.5. Seed sense sequences for SNP KASP assay design. 90
Table 3.6. Effects of repeat length of the expanded allele on age at onset. 90
Table 4.1. 12 genes significantly upregulated in HD blood from Borovecki et al. (2005). 99
Table 4.2.Top 10 up and downregulated genes in HD blood from Mastrokolias et al. (2015). 99
Table 4.3. Track-HD and Leiden cohorts for RNA-Seq analysis. 101
Table 4.4. Top 10 genes from the differential expression analysis in the combined Track-HD and Leiden cohort. 105
Table 4.5. Overlap analysis of Track-HD and Leiden cohorts shows that a significant excess of pathways are associated with HD ($p < 0.05$) in both datasets. 106
Table 4.6. The 10 most significantly up and downregulated 'generic' pathways in HD versus control blood GSEA. 106
Table 4.7. Top genes in top pathways. 108
Table 4.8. Groups of pathways upregulated in HD blood vs controls. 109
Table 4.9. Groups of pathways downregulated in HD blood vs controls. 110
Table 4.10. Overlap between HD blood and myeloid cells. 110
Table 4.11. Top 10 upregulated pathways that overlap between HD blood and myeloid cells. 111
Table 4.12. WGCNA brain expression modules in HD versus control blood. 112
Table 4.13. Top 10 genes in WGCNA modules. 113
Table 4.14. Brain expression modules significantly dysregulated both in HD brain and HD blood. 114
Table 4.15. Top 10 genes in module 48 (CNpos2) that are dysregulated ($p < 0.05$) in both blood and caudate, ranked by their kME value. 114

Table 4.16. Top 10 pathways dysregulated (p < 0.05) in both HD prefrontal cortex (Labadorf et al., 2015) and blood. . 115
Table 4.17. Top 10 modules dysregulated (p < 0.05) in both HD prefrontal cortex (Labadorf et al., 2015) and blood. .. 116
Table 4.18. Top 10 genes with expression in correlation with disease severity (total motor score). 116
Table 4.19. Top 10 pathways enriched for up and downregulation in HD blood that also enriched for genes correlated with disease severity (TMS) in the same direction. 117
Table 4.20. Modules dysregulated in HD blood that also correlated with disease severity (TMS) in the same direction. 118
Table 4.21. Top 10 differentially expressed genes from Mastrokolias et al (Mastrokolias et al., 2015) that correlated with disease severity (TMS) in Track-HD blood. 119
Table 4.22. Modules from Gibbs et al. (2010) that are dysregulated in both Alzheimer's disease brain (International Genomics of Alzheimer's Disease, 2015) and HD blood. 119
Table 4.23. Top 10 co-expression modules from Alzheimer's disease brain(Zhang et al., 2013) that are dysregulated in HD blood. 120
Table 5.1. Antibodies for immunofluorescence. 131
Table 5.2. Taqman qPCR probes. 132
Table 5.3. Repeat expansion analysis in ReN VM 129Q NSCs chronically stressed with H_2O_2 during differentiation for 48 days. 135
Table 5.4. Repeat expansion analysis in ReN VM 129Q NSCs 44d after initiation of differentiation, chronically stressed with H_2O_2 from day 15. 136
Table 5.5. CAG repeat sizing analysis for two single cell clones in culture and during differentiation. 139
Table 5.6. Exponential modelling of modal CAG expansion in 109Q iPSCs. 151
Table 6.1. Full length Myc-tagged FAN1 construct. 164
Table 6.2. Sequencing primers to confirm SDM. 166
Table 7.1. Animals used in the toxicity study. 199
Table 7.2. Animals used in experimental study. 200
Table 7.3. TaqMan qPCR probes. 202
Table 7.4. Change in modal CAG relative to 12 day tail. 227
Table 7.5. Somatic instability index relative to 12d tail. 229
Table 7.6. Proportional expansion analysis. 231
Table 8.1. Regression models of the relationships between allele structures, relative rate of somatic expansion and disease phenotypes in Huntington's disease and myotonic dystrophy type 1. 244
Table 8.2. MSH3 9 bp tandem repeat alleles observed in HD and DM1 cohorts. 246
Table 8.3. MSH3 9 bp tandem repeat alleles and their association with phenotypes in DM1 and HD. 250
Table 8.4. Detailed investigation of the role of repeat alleles in phenotypic modification. 251
Table 8.5. MSH3 exon 1 region variants. 254
Table 8.6. MSH3 exon 1 region variants and the association of their alternative alleles with phenotypes in DM1 and HD. 255
Table 8.7. Associations of SNP alternative alleles with phenotypes conditional on the repeat structure (Table 4). 256

Table 8.8. MSH3 exon 1 region haplotypes. 257
Table 8.9. MSH3 exon 1 region haplotypes and their association with phenotypes in DM1 and HD. 258
Table 10.1. Nextera XT Index Kit v2 primers for the MSH3 repeat region. 283
Table 10.2. Base-wise conservation scores across the MSH3 exon 1 9bp tandem repeat region. 307
Table 10.3. MSH3 and DHFR expression quantitative trait loci associated with phenotypes in HD and DM1. 309
Table 10.4. Transcriptome-wide association study (TWAS) of HD prefrontal cortex. 313

Abbreviations

6-MTG	6-methylthioguanine
AAO	Age at onset
AAV	Adeno-associated virus
ABI	Applied Biosystems
ACMG	American College of Medical Genetics
ACTB	Actin, cytoplasmic 1
AD	Alzheimer's disease
ADORA2A	Adenosine A2a Receptor
ADP	Adenosine diphosphate
ADPRH	ADP-Ribosylarginine Hydrolase
ADR	Adrenal gland
ALS	Amyotrophic lateral sclerosis
APTX	Aprataxin
ARPP21	CAMP Regulated Phosphoprotein 21
AT	Ataxia telangiectasia
ATLD	Ataxia-telangiectasia-like disorder
ATM	Serine-protein kinase ATM
ATP	Adenosine triphosphate
BDNF	Brain-derived neurotrophic factor
BER	Base excision repair
BMI	Body mass index
BS	Brainstem
BSA	Bovine serum albumin
BWA	Burrows-Wheeler Aligner
BWT	Burrows-Wheeler transform
CAG	Cytosine-Adenine-Guanine
CALB1	Calbindin 1
CB	Cerebellum
CBM	Carbamazepine
CBP	Phosphoprotein associated with glycosphingolipid-enriched microdomains 1
CHDI	Cure Huntington's Disease Initiative
CJD	Creutzfeldt-Jakob disease
CMC	CommonMind Consortium
CN	Caudate nucleus
CNS	Central nervous system
CNV	Copy-number variation

CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSF	Cerebrospinal fluid
CTD	C-terminal domain
CTG	Cytosine-Thymine-Guanine
CTIP	DNA endonuclease RBBP8
CX	Cortex
DAPI	4’,6-diamidino-2-phenylindole
DARPP-32	Protein phosphatase 1 regulatory subunit 1B (PPP1R1B)
DDR	DNA damage response
DHFR	Dihydrofolate reductase
DLX1	Distal-Less Homeobox 1
DLX2	Distal-Less Homeobox 2
DLX5	Distal-Less Homeobox 5
DLX6	Distal-Less Homeobox 6
DMEM	Dulbecco's Modified Eagle's medium
DMPK	Myotonin-protein kinase
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DRD1	Dopamine Receptor D1
DRD2	Dopamine Receptor D2
DRIP	DNA-RNA immunoprecipitation
DRPLA	Dentatorubral-pallidoluysian atrophy
DSB	Double strand break
DSBR	Double strand break repair
DTT	Dithiothreitol
EBF1	Early B Cell Factor 1
EBV	Epstein-Barr virus
ECACC	European Collection of Authenticated Cell Cultures
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylene glycol-bis(β-aminoethyl ether)-N,N,N’,N’-tetraacetic acid)
EHDN	European Huntington's Disease Network
EMBL	European Molecular Biology Laboratory
EMQN	European Molecular Genetic Quality Network
EMS	Ethylmethanesulphonate
EPC	Erythroid progenitor cells
FA	Fanconi anaemia
FACS	Fluorescence activated cell sorting
FAN1	FANCD2 And FANCI Associated Nuclease 1

FANCA	Fanconi anemia group A protein
FANCI	Fanconi anemia group I protein
FANCM	Fanconi anemia group M protein
FB	Fibroblast
FBS	Fetal bovine serum
FC	Frontal cortex
FCTX	Frontal cortex
FDR	False discovery rate
FGF	Fibroblast growth factor
FOXP2	Forkhead Box P2
FRDA	Friedreich ataxia
FRT	Flippase recognition target
FTD	Frontotemporal dementia
FTL	Ferritin light chain
FXN	Frataxin
FXS	Fragile X syndrome
GAD1	Glutamate Decarboxylase 1
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GB	Gillian Bates
GDNF	Glial cell line-derived neurotrophic factor
GFP	Green fluorescent protein
GSEA	Gene Set Enrichment Analysis
GSX2	GS Homeobox 2
GWAS	Genome-wide association study
HBV	Hepatitis B virus
HCL	Hydrochloric acid
HCV	Hepatitis C virus
HD	Huntington's disease
HDAC	Histone deacetylase
HDL-1	Huntington’s disease like syndrome 1
HDL-2	Huntington’s disease like syndrome 2
HGVS	Human Genome Variation Society
HIPP	Hippocampus
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HNPCC	Hereditary nonpolyposis colorectal cancer
HTT	Sodium-dependent serotonin transporter
HWE	Hardy-Weinberg equilibrium

ICL	Interstrand crosslink
IDCL	Interdomain connector loop
IDL	Insertion-deletion loops
IGAP	International Genomics of Alzheimer's Disease Consortium
IGFALS	Insulin Like Growth Factor Binding Protein Acid Labile Subunit
IHC	Immunohistochemistry
IP	Immunoprecipitation
IRB	Institutional review board
IRES	Internal ribosome entry site
IS	Intrastriatal
IT	Intrathecal
IV	Intravenous
KD	Knockdown
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIN	Karyomegalic interstitial nephritis
LB	Lymphoblastoid cell
LDH	Lactate dehydrogenase
LGE	Lateral ganglionic eminence
LHX6	LIM Homeobox 6
LIV	Liver
LOAD	Late-onset Alzheimer's disease
MAPK	Mitogen-activated protein kinase
MEF	Mouse embryonic fibroblast
MEM	Minimum Essential Medium Eagle
MGI	Mouse Genome Informatics
MHC	Major histocompatibility complex
MHF	Centromere Protein S
MJD	Machado–Joseph disease
MMC	Mitomycin C
MMLV	Moloney Murine Leukemia Virus Reverse Transcriptase
MMR	Mismatch repair
MMS	Methyl methanesulfonate
MRC	Medical Research Council
MRC PPU	Medical Research Council Protein Phosphorylation and Ubiquitylation Unit
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
MSI	Microsatellite instability
MSN	Medium spiny neuron

MTM-HD	Multiple Tissue Molecular Signatures in Huntington's Disease
MTT	3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide
NAD	Nicotinamide adenine dinucleotide
NADPH	Nicotinamide adenine dinucleotide phosphate
NANOG	Nanog Homeobox
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NEB	New England Biolabs
NEFL	Neurofilament light polypeptide
NER	Nucleotide excision repair
NES	Normalised effect size
NGS	Next-generation sequencing
NHEJ	Nonhomologous end joining
NKX2-1	NK2 Homeobox 1
NOLZ1	Zinc Finger Protein 503
NSC	Neural stem cell
NTD	N-terminal domain
NVC	Naive Variant Caller
OB	Olfactory bulb
OCP	Oral contraceptive pill
OMIM	Online Mendelian Inheritance in Man
OPL	Outer plexiform layer
ORF	Open reading frame
PACRGL	Parkin Coregulated Like
PANDAS	Paediatric autoimmune neuropsychiatric disorders associated with streptococcal infection
PBL	Peripheral blood lymphocytes
PBMC	Peripheral blood mononuclear cells
PBS	Phosphate-buffered saline
PCA	Principal component analysis
PCNA	Proliferating cell nuclear antigen
PCNP	PEST proteolytic signal-containing nuclear protein
PCR	Polymerase chain reaction
PCTP	Phosphatidylcholine Transfer Protein
PD	Parkinson's disease
PDGFD	Platelet Derived Growth Factor D
PDPN	Podoplanin
PENK	Proenkephalin
PEST	PEST Proteolytic Signal Containing Nuclear Protein

PFA	Paraformaldehyde
PFC	Prefrontal cortex
PGC1α	PPARG Coactivator 1 Alpha
PHAROS	Prospective Huntington At Risk Observational Study
PHE	Public Health England
PIGH	Phosphatidylinositol Glycan Anchor Biosynthesis Class H
PIGN	Phosphatidylinositol Glycan Anchor Biosynthesis Class N
PIGX	Phosphatidylinositol Glycan Anchor Biosynthesis Class X
PIP box	PCNA-interacting peptide (PIP) box
PMS2	PMS1 Homolog 2, Mismatch Repair System Component
PNPK	Polynucleotide kinase 3'phosphatase
PRNP	Major prion protein
PTN	Phenytoin
QC	Quality control
RAN translation	Repeat-associated non-ATG (RAN) translation
REST	RE1-silencing transcription factor
RIPA	Radioimmunoprecipitation assay buffer
RMA	Robust Multi-array Average
RNA	Ribonucleic acid
ROB	Rest of brain
RPA	Replication protein A
RPMI	Roswell Park Memorial Institute medium
RT-qPCR	Reverse transcription quantitative polymerase chain reaction
SAGE	Serial analysis of gene expression
SAM	Sequence Alignment Map
SAP	SAF-A/B, Acinus and PIAS
SBMA	Spinal and bulbar muscular atrophy
SBS	Sequencing by Synthesis
SCA	Spinocerebellar ataxia
SCC	Single cell clone
SD	Standard deviation
SDHA	Succinate Dehydrogenase Complex Flavoprotein Subunit A
SDM	Site-directed mutagenesis
SDS-PAGE	Sodium dodecyl sulphate-polyacrylamide gel electrophoresis
SE	Standard error
SEM	Standard error of the mean
SIFT	Sorting Intolerant From Tolerant
SII	Somatic instability index

SIN	Somatic instability network
SIX3	SIX Homeobox 3
SL	Left striatum
SLE	Systemic lupus erythematosus
SMAD	Caenorhabditis elegans SMA ("small" worm phenotype) and Drosophila MAD ("Mothers Against Decapentaplegic") family of genes
SNP	Single nucleotide polymorphism
SPATAX	Spastic paraplegias (SP) and cerebellar ataxias (CA) network
SPSS	Statistical Package for the Social Sciences
SSBR	Single-strand break repair
SSRI	Selective serotonin reuptake inhibitor
STR	Striatum
STRING	Database of known and predicted protein-protein interactions
SVP	Sodium valproate
TAC1	Tachykinin Precursor 1
TATA	Promoter region containing repeating T and A base pairs
TBP	TATA-box-binding protein
TBS	Tris-buffered saline
TCTX	Temporal cortex
TDP1	Tyrosyl-DNA phosphodiesterase 1
TEMED	Tetramethylethylenediamine
TFC	Total functional capacity
TLS	Translesion synthesis
TMS	Total motor score
TNF	Tumor necrosis factor
TRIS	Tris(hydroxymethyl)aminomethane
TWA	Transcriptome-wide association
TWAS	Transcriptome-wide association study
TX-100	Triton X-100
UBZ	Ubiquitin-binding zinc finger
UCL	University College London
UCLH	University College London Hospitals
UCSC	University of California Santa Cruz
UHDRS	Unified Huntington's Disease Rating Scale
UK	United Kingdom
UTR	Untranslated region
UV	Ultraviolet
VC	Visual cortex

VCF	Variant Call Format
VCP	Valosin Containing Protein
VM	Ventral mesencephalon
VRR Nuc	Viral replication and repair nuclease domain
WES	Whole exome sequencing
WGCNA	Weighted gene correlation network analysis
WGE	Whole ganglionic eminence
WPRE	Woodchuck Hepatitis Virus (WHP) Posttranscriptional Regulatory Element
WT	Wild type
XL	X-linked inheritance
XLD	X-linked dominant inheritance
XPF	DNA repair endonuclease XPF

Chapter 1 Introduction

1.1 DNA repeat expansion

1.1.1 Repetitive DNA

Over 65% of the human genome consists repetitive elements from microsatellites of a few base pairs up to arrays of whole genes, which have a range of functions, including the regulation of chromatin structure and transcription (Hall et al., 2017, Budworth and McMurray, 2013, Biscotti et al., 2015). Microsatellites are common, constituting around 2% of the genome, so it is not their presence itself that is pathogenic.

1.1.2 Repeat expansion diseases

For reasons which remain unclear, repeat expansion diseases often have a neurological phenotype (Madabhushi et al., 2014, Neil et al., 2017). Huntington's disease (HD), myotonic dystrophy type 1 (DM1), spinal and bulbar muscular atrophy (SBMA), dentatorubral-pallidoluysian atrophy (DRPLA) and several spinocerebellar ataxias (SCA 1,2,3,6,7,12 and 17) are caused by (CAG)n/(CTG)n repeats, Friedreich's ataxia (FA) by (GAA)n, fragile X syndrome (FXS) by (CGG)m, myotonic dystrophy type 2 (DM2) by (CCTG)n, SCA10 by (ATTCT)n and C9orf72 by (GGGGCC)n (Neil et al., 2017). For the polyglutamine diseases, the pathogenic threshold is around 35-45 CAG repeats (Massey and Jones, 2018). Pathogenic repeat expansions can occur either outside or within the coding sequence, with non-coding expansions tending to be longer. Though mechanisms driving repeat expansion may be similar, the repeats occur in different genomic contexts and proteins which are functionally unrelated. It is likely that differences in protein function and expression profile produce the distinctive phenotype of each condition (Orr and Zoghbi, 2007).

Trinucleotide repeat diseases are individually rare, but together represent a relatively common group of neurodegenerative diseases and a significant source of morbidity. Fragile X syndrome is caused by a CGG expansion in the *FMR1* gene, and is the most common, affecting 1/4000 males and 1/8000 females. Myotonic dystrophy (DM1), caused by a CTG expansion in *DMPK*, affects around 1/8,000 and Huntington's disease, caused by a CAG expansion in *HTT*, affects around 1/10,000. The spinocerebellar ataxias each have a prevalence of around 1/100,000 (McKusick, 2007).

1.1.3 Repeat instability

1.1.3.1 Somatic instability

Pathogenic DNA repeats are inherently unstable and tend to expand throughout life in particular tissues, depending on the disease. The extent to which somatic expansion influences human disease course is not known, but it is seen in postmortem human HD brain neurons and correlates with earlier onset (Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009). Though somatic expansions in HD brain tissue may be large (Kennedy et al., 2003), the level of somatic mosaicism in peripheral tissues, such as blood, is low (Telenius et al., 1995, Leeflang et al., 1995). Despite this, using a single molecule PCR approach, length-dependent, expansion biased somatic mosaicism has previously been demonstrated in HD patient buccal cells (Veitch et al., 2007). In transgenic and knock-in mouse models there is expansion in postmitotic neurons of the brain, particularly the striatum which correlates with symptom onset (Mangiarini et al., 1997) and may explain its selective vulnerability (Gonitel et al., 2008, Lee et al., 2011a). There is also expansion in the liver, but stability in the cerebellum, blood and tail. In DM1, large expansions occur in muscle, the tissue most prominently affected, which may cease after terminal differentiation (Thornton et al., 1994, Zatz et al., 1995), as well as in lymphocytes

throughout life (Martorell et al., 1995). In HD (Benitez et al., 1995) and SBMA (Jedele et al., 1998), instability is detectable only in adults, and not foetuses, whereas in fragile X syndrome it is seen only in foetal tissue and not postnatally (Reyniers et al., 1999, Devys et al., 1992, Taylor et al., 1999). Unlike other polyglutamine diseases, the SBMA CAG tract is stable in CNS and, like DM1, expands in muscle (Tanaka et al., 1999). The observation of repeat instability in postmitotic CNS neurons (Gonitel et al., 2008) and continued expansion when the cell cycle is arrested in transgenic mouse cells (Gomes-Pereira et al., 2014b) suggests expansion, at least of this type of repeat, occurs during DNA repair or transcription, rather than replication.

The tissue specificity of somatic instability and neuropathology often overlap, for example in HD, DM1 and FRDA (Goula et al., 2012). As discussed above, in HD the striatum shows the most prominent CAG expansion and degeneration, but both also occur in the cortex and are limited in the cerebellum (Wheeler et al., 1999, Goula et al., 2012, Shelbourne et al., 2007a). In DM1, the striatum has not been studied, but CTG expansion is greatest in muscle and also occurs in cortex, though is limited in cerebellum (Anvret et al., 1993, Wong et al., 1995, Ashizawa et al., 1993, Lopez Castel et al., 2011). In Friedreich's ataxia (FRDA), the GAA repeat expands significantly in the cerebellum and dorsal root ganglia, two tissues conspicuously affected by the disease (De Biase et al., 2007, Clark et al., 2007b). However, the CAG expansion profile in SCA1 (Watase et al., 2003, Kraus-Perrotta and Lagalwar, 2016, Zuhlke et al., 1997, Lopes-Cendes et al., 1996), SCA3 (La Spada, 1997, Hashida et al., 1997) and DRPLA (Hashida et al., 2001, Watanabe et al., 2000, Aoki et al., 1996, Zuhlke et al., 1997, Lopes-Cendes et al., 1996, Takano et al., 1996, Ueno et al., 1995) is similar to that of HD, with instability in basal ganglia and cortex, and stability in cerebellum, a prominently affected tissue (La Spada, 1997). Therefore, there is not always a clear correlation between somatic instability and tissue vulnerability. Brain region and cell type-specific instability may reflect the different developmental history of these regions, or tissue and cell specific factors such as DNA repair protein expression.

Interruptions in the repeat sequence, which reduce the stability of hairpin structures, have been shown to restrict expansion in many trinucleotide repeat disorders, including the HD, SCAs 1-3 and 17, fragile X syndrome, Friedreich's ataxia and DM1 (Massey and Jones, 2018), and delay onset in HD (Lee et al., 2019, Wright et al., 2019). CAG interruptions usually alter the third base of the codon and can reduce hairpin loop formation (Menon et al., 2013, Pearson et al., 1998, Sobczak and Krzyzosiak, 2004, Kraus-Perrotta and Lagalwar, 2016).

1.1.3.2 Germline instability

In all trinucleotide repeat disorders the repeat is also unstable in germ cells, causing the length to increase in successive generations (Jones et al., 2017). Most polyglutamine disorders have a paternal expansion bias (Pearson et al., 2005b), and there is significant CAG length mosaicism in HD patient sperm that correlates with expansion on transmission (Telenius et al., 1995). In transgenic mice, expansion occurs after meiosis, again implicating DNA repair or transcription rather than replication (Kovtun and McMurray, 2001). Expansions increase with increasing paternal age in several transgenic CAG/CTG mouse models (Pearson et al., 2005b). Contrastingly, there is a maternal expansion bias in fragile X syndrome and DM1, and there is a paternal contraction bias in SCA8, Friedreich's ataxia and fragile X syndrome, which may be related to reduced methylation of repeats in testes (Pearson et al., 2005b). Unlike male germ cells, oogenic meiosis occurs *in utero* and then arrests for years until puberty, resuming minutes before ovulation and continuing until fertilisation (Pearson et al., 2005b). Therefore, the relationship between instability and cell division, transcription and

DNA repair is not straightforward, and it is likely that tissue-specific and *cis* or *trans*acting factors act to modify expansion (Pearson et al., 2005b).

1.1.3.3 Pathogenicity

Though many unstable repeat regions have been linked to disease, the mechanisms by which their expansion above a threshold lead to disease remains unclear. In fragile X syndrome and Friedreich's ataxia, repeat expansion results in silencing of expression (Colak et al., 2014). In myotonic dystrophy (DM1), the expansion causes the formation of RNA foci (Thornton, 2014). Repeat-associated non-ATG translation (RAN) was first identified in DM1 and SCA8 (Zu et al., 2011) and has also been found in other trinucleotide disorders including Huntington's disease and fragile X syndrome (Banez-Coronel et al., 2015, Cleary and Ranum, 2014). Toxicity of the resulting dipeptides has been demonstrated in C9ORF72 associated frontotemporal dementia and amyotrophic lateral sclerosis, though their role in other repeat expansion diseases is unclear. In the CAG expansion diseases the repeat-containing protein aggregates as insoluble protein inclusions within cells, a feature that is also seen in other neurodegenerative diseases such as Alzheimer's (Knowles et al., 2014).

1.2 Huntington's disease

Huntington's disease (HD), the most common monogenic neurodegenerative disorder in the developed world (Evans et al., 2013), is caused by a CAG repeat expansion in the *HTT* gene and is characterised by motor, cognitive and psychiatric features. It was named after George Huntington, who described the condition in 1872 (Huntington, 1872), but it was not until 1983 that the genetic locus was mapped (Gusella et al., 1983) and 1993 when the gene was discovered (Group, 1993). Onset occurs around 45 years on average and inversely correlates with CAG repeat length (Langbehn et al., 2010). The disease progresses inexorably and, with the exception of late-onset cases, is uniformly fatal a median of 18 years from motor onset (Ross et al., 2014). HD is currently incurable and no treatments slow progression.

1.2.1 Epidemiology

In the UK, HD affects around 1 in every 7300 people (Bates et al., 2015c). Prevalence has progressively increased owing to increasing survival and the introduction of a genetic test that has allowed the diagnosis of *de novo* and late onset cases (Evans et al., 2013, Morrison, 2012). HD is found around the world, but at higher frequencies in populations of European descent (Bates et al., 2015c). In East Asian populations the prevalence is around 1-7 per million, potentially because mean CAG repeat length in the population is shorter.

1.2.2 Aetiology

HD is caused by a (CAG)n repeat expansion in exon 1 of the *HTT* gene on chromosome 4 (Group, 1993). Repeat lengths from 36 to 39 units have reduced penetrance, and those ≥40 are fully penetrant (Bates et al., 2015c). The mechanism underlying this length-dependent trigger remains unclear, as does the nature of the toxic species underlying its pathogenicity (Bates et al., 2015c).

1.2.3 Pathogenesis

HTT is expressed throughout the body, though at varying levels in different cell types. The protein is mostly cytoplasmic, but forms can be found in the nucleus and cytoplasm, and are able to shuttle between the two compartments (Bates et al., 2015c). It has many interacting partners, particularly at the N-terminus, suggesting it acts as a scaffold for complexes

of proteins (Ross and Tabrizi, 2011). Its normal function is still unclear, but it has roles in nervous system development and protein homeostasis.

Expansion of the *HTT* CAG repeat results in neuronal dysfunction and death through numerous mechanisms. It undergoes extensive post-translational modification, with proteolytic fragmentation producing a toxic N-terminal fragment of around 100 amino acids that readily aggregates (Bates et al., 2015c). Aggregated inclusions rich in mutant HTT (mHTT) form in neuronal nuclei, but also elsewhere in the cell, including the cytoplasm, dendrites and axon terminals (Vonsattel, 2008). However, several reports have suggested their density does not correlate with cell toxicity, leading to the idea they may be a protective cellular response to misfolded protein (Kim et al., 1999, Arrasate et al., 2004, DiFiglia et al., 2007). Cells may be able to take up small fibrils of polyglutamine protein, which seed aggregates by recruiting endogenous protein in a prion-like mechanism of cell-to-cell transmission (Cicchetti et al., 2014, Pecho-Vrieseling et al., 2014). Expression of mHTT causes proteostasis to deteriorate, with chaperone levels decreasing, endoplasmic stress increasing and the proteasomal and autophagy systems becoming compromised (Bates et al., 2015c), limiting cells' ability to respond to stress. Toxic forms of mutant huntingtin disrupt many fundamental cellular processes, including transcription (Seredenina and Luthi-Carter, 2012), mitochondrial function (Reddy and Shirendeb, 2012, Johri et al., 2013), synapses (Nithianantharajah and Hannan, 2013) and intracellular signalling (Labbadia and Morimoto, 2013), cellular transport (Reddy and Shirendeb, 2012) and secretion (Vidal et al., 2011), endocytic recycling (Kim et al., 1999), and the immune system (Ellrichmann et al., 2013). *HTT* RNA itself may have toxic properties, potentially involving antisense mechanisms or toxic repeat associated non-ATG (RAN) translation proteins (Banez-Coronel et al., 2015, Ross and Tabrizi, 2011, Cattaneo et al., 2005). The interaction of these disrupted pathways produces an extremely complex set of pathogenic mechanisms (Ross and Tabrizi, 2011, Bates et al., 2015c).

Medium spiny neurons (MSN) of the striatum are selectively vulnerable. The cause is unclear, but D2 receptors may be a factor as they are expressed by indirect, but not direct pathway MSNs and they have been implicated in pathogenesis (Deyts et al., 2009). Other potential mechanisms include loss of brain derived neurotrophic factor (BDNF) support or glutamate excitotoxicity from cortico-striatal projections (Ross and Tabrizi, 2011).

1.2.4 Pathology

In early disease, the brain can look macroscopically normal, but as disease progresses there is prominent atrophy of the basal ganglia, particularly the caudate nucleus, as well as cortical atrophy with ventricular dilatation (Wood, 2012). Microscopically there is selective loss of MSNs in the striatum (Ferrante et al., 1985) and microglial activation (Sapp et al., 2001). The Vonsattel grade provides a histopathological classification in symptomatic patients ranging from 0, with no gross or microscopic abnormalities, to 4, in which there is extreme atrophy (Vonsattel et al., 1985).

HD research has traditionally focused on the brain due to the presence of characteristic mutant huntingtin protein aggregates (Bates et al., 2015c) and because the prominent symptoms and signs can be linked to neurodegeneration in the basal ganglia and cerebral cortex (van der Burg et al., 2009). However, mutant *HTT* is ubiquitously expressed (Trottier et al., 1995) and mounting evidence suggests it has direct effects in peripheral tissues (van der Burg et al., 2009, Carroll et al., 2015). HD patients demonstrate peripheral immune dysfunction presymptomatically (Tai et al., 2007a, Bjorkqvist et al., 2008, Kwan et al., 2012c, Träger et al., 2015), as well as weight loss that leads to cachexia with advancing disease (Carroll et al., 2015). There is progressive muscle wasting (Busse et al., 2008), endocrine dysfunction (Saleh et al., 2009),

liver impairment (Carroll et al., 2015) and cardiac dysfunction (Lanska et al., 1988, Mihm et al., 2007, Pattison et al., 2008). Mutant HTT protein aggregates can be found in the peripheral tissues of HD mice (Orth et al., 2003), as well as advanced patients (Turner et al., 2007). These peripheral features may contribute to central nervous system (CNS) pathology, disease progression and mortality (Carroll et al., 2015, van der Burg et al., 2009), and strongly suggest that HD is a systemic disorder. It is unclear whether peripheral effects are distinct, or parallel those in the brain. Mechanisms of dysfunction include transcriptional dysregulation, disordered protein folding, deficient protein degradation and inflammatory activation (Bates et al., 2014, Bates et al., 2015c).

1.2.5 Clinical features

After an asymptomatic premanifest period, a prodromal phase with subtle motor, cognitive and psychiatric features often precedes formal diagnosis of motor onset by up to 15 years (Bates et al., 2015c). Motor onset occurs at around 45 years on average (Langbehn et al., 2010) and is followed by inexorable progression (Ross et al., 2014). Onset is often difficult to clearly discern, and many have early psychiatric and cognitive symptoms. Definitive diagnosis is made when there are unequivocal motor signs.

1.2.5.1 Motor

Motor manifestations often begin with subtle restlessness, fidgeting and fine involuntary movements, and progress to chorea. Eye movements are an early sign, with delayed and slow saccades and impaired pursuit with saccadic intrusions (Wood, 2012). There are also varying degrees of dystonia, parkinsonism and bradykinesia, but impairment of voluntary motor function is often more functionally disabling. Impaired walking and postural reflexes lead to falls. Dysarthria causes communication problems, with much frustration for patients and carers, and as the disease progresses patients often become mute. Dysphagia is common, and choking is often reported early. Initially it can be the result of impulsive and disordered eating, but later there is mechanical discoordination.

1.2.5.2 Cognition

Cognitive and psychiatric features are usually the most disabling. Cognitive impairment is universal, though it affects specific functions so the term ‘dementia’ tends not to be used. There is limited impact on language and spatial skills, but prominent involvement of executive function, with impaired planning, judgement and multi-tasking. There tends to be psychomotor slowing, and apathy and a lack of initiative can make caring challenging. Patients themselves often complain of poor concentration and attention. With progression, patients are less able to care for themselves, though often lack insight (Wood, 2012).

1.2.5.3 Psychiatric

Depression and anxiety are common (Craufurd and Snowden, 2002) and suicide rates are higher than in the general population (Farrer, 1986). Irritability is a common feature, and some can be aggressive. Obsessions and compulsions can develop. Psychosis is rare.

1.2.5.4 Systemic features

Patients often lose weight. The cause is thought to be multifactorial, including poor intake, dysphagia and increased energy expenditure due to involuntary movements. However, *HTT* is expressed ubiquitously (Trottier et al., 1995) and the peripheral phenotype is well established (van der Burg et al., 2009, Carroll et al., 2015, Tai et al., 2007a, Bjorkqvist et

al., 2008, Kwan et al., 2012c, Träger et al., 2015, Busse et al., 2008, Saleh et al., 2009, Lanska et al., 1988, Mihm et al., 2007, Pattison et al., 2008, Orth et al., 2003, Turner et al., 2007). Higher premorbid BMI is associated with slower progression, so patients are encouraged to maintain their weight (Myers et al., 1991). Patients often have a disturbed sleep-wake cycle due to disruption of circadian rhythm (Morton, 2013).

1.2.5.5 Juvenile-onset HD

This is defined as onset before 20 years and is usually associated with over 60 CAG repeats (Fusilli et al., 2018). The disease is more severe and carries a shorter life expectancy. Patients tend to have a more akinetic-rigid form with minimal chorea but increased dystonia, as well as seizures.

1.2.6 Diagnosis

Genetic testing is definitive, whereas imaging, blood and cerebrospinal fluid analysis are not particularly useful in diagnosis. MRI may show caudate and cortical atrophy early in disease (Wood, 2012, McColgan et al., 2015, Gregory et al., 2018, Tabrizi et al., 2011b).

Around 1% of those presenting with HD signs test negative for the *HTT* expansion (Andrew et al., 1994a). The differential diagnosis of autosomal dominant HD phenocopies is broad and is summarised in the tables below. Ultimately, a genetic diagnosis is found in only 3% of this subpopulation, the commonest being C9orf72 in 1.9-5% (Beck et al., 2013).

Gene (mutation)	Disease	Inheritance	Pointers	Frequency
C9orf72 (GGGGCC repeat)	ALS/FTD	AD		1.9-5%
JPH3 (CTG/CAG repeat)	HDL-2	AD	African or Middle Eastern	1.3-4.5%
VPS13A (chorein)	Choreoacanthocytosis	AR	Acanthocytes, orofacial dyskinesia with tongue protrusion, lip biting	0.4-3%
Mutations in mitochondrial DNA and nuclear DNA encoding mitochondrial proteins	Mitochondrial disease	-	Myoclonus, dementia, muscle biopsy (ragged red fibres)	1.9%
TBP (CAG/CAA repeat)	SCA17	AD		0.5-1.8%
FXN (GAA repeat)	Friedreich's ataxia	AR	Ataxia	0.4-1%
CACNA1A	SCA6, Episodic ataxia 2	AD		0.9%
UBQLN2	ALS/FTD	XLD		0.4%
VCP	ALS/FTD	AD		0.4%
PRNP (octapeptide rpt)	HDL-1	AD		0.4%

Table 1.1 Genetic HD phenocopies.

AD – autosomal dominant, AR – autosomal recessive, ATN1 – atrophin 1, DRPLA – Dentatorubral-pallidoluysian atrophy, FTL – ferritin light chain, HDL-1/2 – Huntington's disease like syndrome 1/2, JPH3 – Junctophilin 3, PRNP – prion protein, SCA – spinocerebellar ataxia, TBP – TATA box-binding protein, (Mariani et al., 2016, Wild et al., 2008, Wild and Tabrizi, 2007a, Wild and Tabrizi, 2007b).

Group	Disease	Frequency
Metabolic	B12 deficiency	0.4%
Immune	Systemic lupus erythematosus (SLE), antiphospholipid syndrome	
Vascular	Basal ganglia stroke	
Infection	AIDS-related	
Post-infectious	Sydenham's chorea, PANDAS (paediatric autoimmune neuropsychiatric disorders associated with streptococcal infection)	
Drugs	Dopamine antagonists (neuroleptics, antiemetics), antiepileptics (phenytoin, carbamazepine, sodium valproate, gabapentin), benzodiazepines, oral contraceptive pill	
Cancer	Paraneoplastic, basal ganglia metastases	

Table 1.2. Acquired HD phenocopies.

(Mariani et al., 2016, Wild et al., 2008, Wild and Tabrizi, 2007a, Wild and Tabrizi, 2007b).

1.2.7 Therapy

Though many therapeutic targets have been identified, none have yet delivered treatments capable of modifying disease course in humans (Bates et al., 2015a, Hughes, 2014). However, many of symptoms are eminently treatable.

1.2.7.1 Motor

Chorea is rarely the most disabling feature and many patients are unaware of its severity. If functionally restrictive, antichoreic medication is used sparingly as none are particularly effective and all can cause side effects. Sulpiride, olanzapine, risperidone and tetrabenazine are options, though the latter in particular carries a risk of depression. Later in the disease, chorea lessens and patients become more rigid and dystonic, at which point antispasticity drugs such as baclofen and clonazepam can be useful. Physiotherapy and walking aids can assist impaired voluntary movement and gait. Early referral to speech and language therapists is useful as exercises and communication aids can help dysarthria, and manoeuvres and modified diets can minimise aspiration from dysphagia. Levodopa may benefit juvenile-onset patients who have prominent parkinsonism (Wood, 2012).

1.2.7.2 Psychiatric

Current practice is largely anecdotal. Depression can be effectively treated with selective serotonin reuptake inhibitors (SSRI) such as citalopram and mirtazapine or cognitive behavioural therapy. Severe anxiety, aggression and impulsive behaviour may respond to newer antipsychotics, including risperidone, olanzapine and quetiapine.

1.2.7.3 Cognition

Patients are less able to care for themselves as the disease progresses, but a common problem is their lack of insight. Formal psychological, occupational and physical therapy assessment can advise on care.

1.2.7.4 Palliative care

Important issues around percutaneous feeding tubes, treatment of recurrent infections and end of life should be discussed early to allow patients to make informed decisions.

1.2.7.5 Future treatments

There are several potential therapeutic developments on the horizon. Reducing *HTT* expression involves nucleotide based suppression using RNA interference (RNAi) and antisense oligonucleotides (ASO), or transcriptional repression using zinc finger proteins (Wild and Tabrizi, 2014). The first medication trialled, IONIS-HTT_{Rx}, non-selectively suppresses both wild type and mutant *HTT* and is infused directly into the cerebrospinal fluid. I am currently a sub-investigator on the phase 1b/2a clinical trial in patients with early manifest Huntington's disease (Trials, 2016).

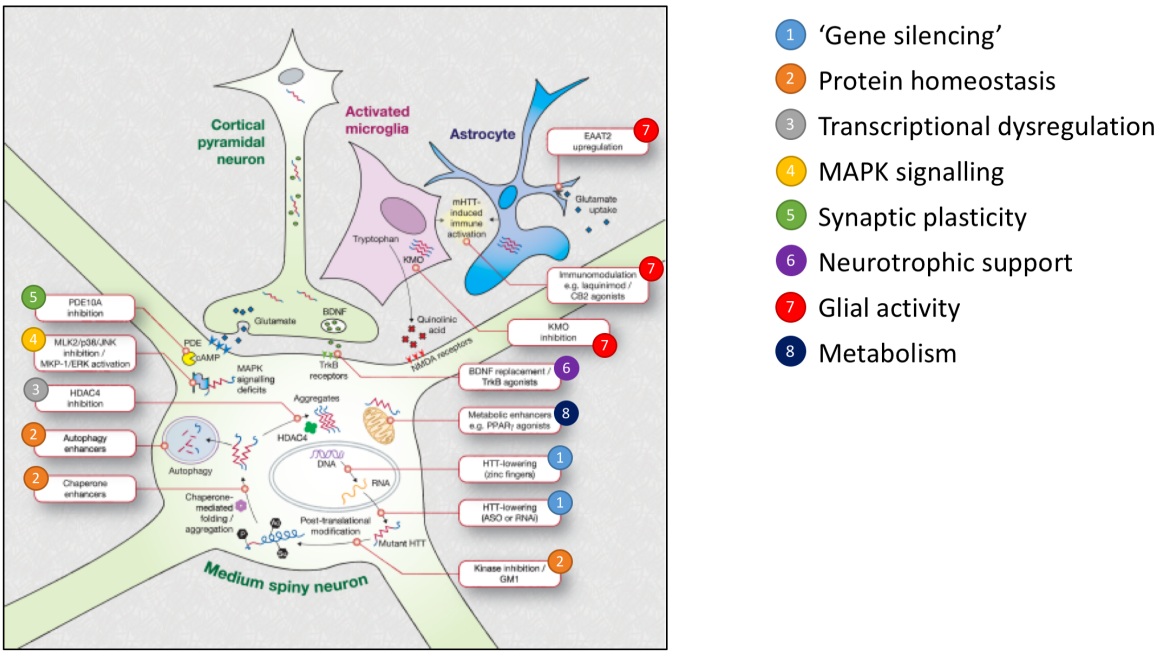


Figure 1.1. Potential future therapeutic targets in Huntington's disease.
The key on the right groups them by target. Adapted from Wild and Tabrizi (2014).

Protein homeostasis targets include kinase inhibitors that modulate HTT phosphorylation (Atwal et al., 2011) and chaperone enhancers to reduce aggregate formation (Labbadia et al., 2012, Sontag et al., 2013). Autophagy enhancing strategies include mTOR inhibition (Renna et al., 2010) and the promotion of mHTT acetylation (Smith et al., 2014, Reilmann et al., 2014). Histone deacetylase inhibitors that regulate chromatin modification are targeting transcriptional dysregulation (Mielcarek et al., 2013). At the synapse, phosphodiesterase inhibitors aim to improve impaired cAMP signalling (Trials, 2015, Beconi et al., 2012) and MAPK signalling inhibitors have been found to be neuroprotective (Taylor et al., 2013, Apostol et al., 2008). BDNF replacement and agonism are being used to address the reduction found in HD brain (Jiang et al., 2013, Conforti et al., 2013, Todd et al., 2014, Simmons et al., 2013). Central and peripheral immune hyperactivity are being targeted with kynurenine 3-monooxygenase (KMO) inhibitors (Zwilling et al., 2011), laquinimod (Comi et al., 2012), cannabinoid receptor agonists (Bouchard et al., 2012c) and excitatory amino-acid transporter 2 (EAAT2) activators (Miller et al., 2008). Trials of antioxidants have so far been ineffective (Mrzljak and Munoz-Sanjuan, 2015), but cellular metabolism is disrupted and modulation of PGC1α has ameliorated mouse models (Jin et al., 2013). Cell replacement has had promising results, but technical and ethical concerns limit its use (Kumar et al., 2016).

1.2.8 Genetic modifiers of Huntington's disease

Motor onset correlates inversely with CAG repeat length, but is still highly variable and can differ by several decades in patients with the same repeat length, as measured in blood (Gusella et al., 2014, Keum et al., 2016). Age at cognitive and psychiatric onset (Keum et al., 2016), as well as age at death (Lanska et al., 1988), also correlate with CAG length, though to a lesser extent (Keum et al., 2016).

The length of the repeat is the main influence on disease course, accounting for around 60% of variation in motor onset (Gusella et al., 2014), but up to 40% of the remaining variation is heritable and due to genetic differences elsewhere in

the genome (Wexler et al., 2004a). Interventions harnessing these mechanisms have the tantalising potential to influence disease course.

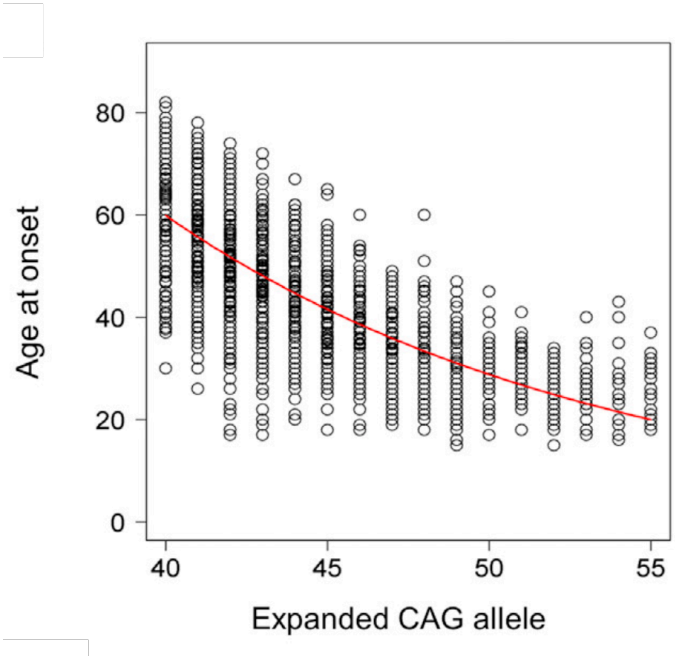


Figure 1.2. Relationship between expanded CAG repeat length (x axis) and onset of diagnostic motor signs (y axis).
Each open circle represents a single HD subject. The red line represents the best fit regression model. Reproduced from GeM-HD (2015).

1.2.8.1 Candidate gene studies

Numerous potential modifiers have been proposed (Djousse et al., 2003, Aziz et al., 2009, Lee et al., 2012d, Lee et al., 2012a, Gusella et al., 2014, Rubinsztein et al., 1997, MacDonald et al., 1999, Cannella et al., 2004, Chattopadhyay et al., 2003, Naze et al., 2002, Zeng et al., 2006, Lee et al., 2012c, Bates et al., 2014, Alberch et al., 2005, Taherzadeh-Fard et al., 2009, Djousse et al., 2004, Li et al., 2006, Gayan et al., 2008). The length of the shorter *HTT* allele was thought to have an effect (Djousse et al., 2003, Aziz et al., 2009), but detailed statistical analysis found the longer allele was fully dominant (Lee et al., 2012d). Sequence variation in *HTT* can be used to define haplotypes, but none were found to significantly modify disease course (Lee et al., 2012a, Gusella et al., 2014). Notable studies of modifiers are listed in the table below, but none withstood statistical analysis. Rubinsztein et al. (1997) reported a TAA repeat polymorphism in the 3' untranslated region of *GRIK2* and this was supported by several subsequent studies (MacDonald et al., 1999, Cannella et al., 2004, Chattopadhyay et al., 2003, Naze et al., 2002, Zeng et al., 2006), but a larger study by Lee et al. (2012c) found no modifier effect. The p.V66M polymorphism in BDNF, a protein known to be functionally relevant in HD, was reported to modify age at onset (AAO) (Alberch et al., 2005), but others have consistently failed to replicate the result (Arning and Epplen, 2013). Taherzadeh-Fard et al. (2009) reported that rs7665116 in *PPARGC1A*, a key regulator of energy metabolism, modified AAO, but this single nucleotide polymorphism (SNP) was later shown to tag Southern European ancestry and patients from these regions are known to have significantly different AAO (Gusella et al., 2014).

Study type	Proposed genetic modifier	Function
HTT variants	Shorter CAG repeat (Djousse et al., 2004), CCG repeat, polymorphisms (Norremolle et al., 2009, Becanovic et al., 2015)	
Candidate gene studies	GRIK2 (Lee et al., 2012b, Rubinsztein et al., 1997)	Glutamate receptor subunit
	UCHL1 (Naze et al., 2002)	Proteasome pathway
	BDNF p.V66M (Alberch et al., 2005)	Neurotrophic factor
	HAP1 p.M441 (Metzger et al., 2008)	Interacts with *HTT*, intracellular trafficking
	PPARGC1A (Taherzadeh-Fard et al., 2009)	PGC-1a, regulator of mitochondrial energy metabolism
	ADORA2A c.C1976T (Dhaenens et al., 2009)	Adenosine receptor
	ATG7 p.V471A (Metzger et al., 2010)	Autophagy
Linkage studies	4p16 (Djousse et al., 2004)	
	6q23-24, 18q22 (Li et al., 2006)	
	2p25, 2q35, 6q22 (Gayan et al., 2008)	
Linkage studies in mice	Mlh1 (Pinto et al., 2013a)	DNA repair

Table 1.3. Notable studies proposing genetic modifiers of Huntington's disease.

1.2.8.2 Linkage studies

Linkage studies in sib-pairs and extended families identified potential modifier loci, though specific genes were not identified (Djousse et al., 2004, Li et al., 2006, Gayan et al., 2008). Linkage studies in the Hdh(Q111) knock in mouse model suggested MMR gene *Mlh1* as a modifier (Pinto et al., 2013a).

1.2.8.3 Genome-wide association studies

The most recent approach to identifying genetic modifiers uses an unbiased genome-wide search. This strategy does not presuppose a pathogenic hypothesis, so findings are data driven. Through the great effort of the HD research community over the last two decades there are now large natural history studies and registries of patients available, along with their DNA samples, permitting the application of this strategy to large populations of HD patients. These include PHAROS (PHAROS, 2006), Predict-HD (Paulsen et al., 2006, Paulsen et al., 2008), TRACK-HD (Tabrizi et al., 2009a), EHDN REGISTRY (Orth et al., 2010, Orth et al., 2011) and COHORT (Dorsey, 2012). Importantly, HD populations may need many fewer subjects than traditional risk studies in order to show associated variation, because in the absence of mutant *HTT* these variants may not produce a selectable phenotype (Gusella et al., 2014). This means genetic modifiers may exist at relatively high frequency because they are not selected against in the general population.

The Genetic Modifiers of Huntington's Disease Consortium (GeM-HD) genome-wide association study (GWAS) of ~6000 HD subjects of European descent identified two loci that influence age of motor onset (GeM-HD, 2015). The experience of the genetics community has emphasised the need for rigorous correction of significance thresholds due to genome-wide testing, meaning a p value of $5x10^{-8}$ is often the criteria for significant association (Sveinbjornsson et al., 2016). At the chromosome 15 locus there were two independent signals; the minor allele at rs146353869 was associated with 6.1 year earlier onset ($p = 4.3x10^{-20}$) and at rs2140734 with 1.4 year later onset ($p = 7.1x10^{-14}$). The second locus was on chromosome 8 and was associated with 1.6 year earlier onset ($p = 2.7x10^{-8}$). The signals replicated in an independent cohort of 3319 European cases. rs35811129, the most significant variant in meta-analysis ($p = 1.55x10^{-22}$) indexes rs2140734. The meta-analysis also identified a genome-wide significant signal on chromosome 3 driven by rs116483964 ($p = 8.84x10^{-9}$) and rs1799977 ($p = 1.19x10^{-8}$).

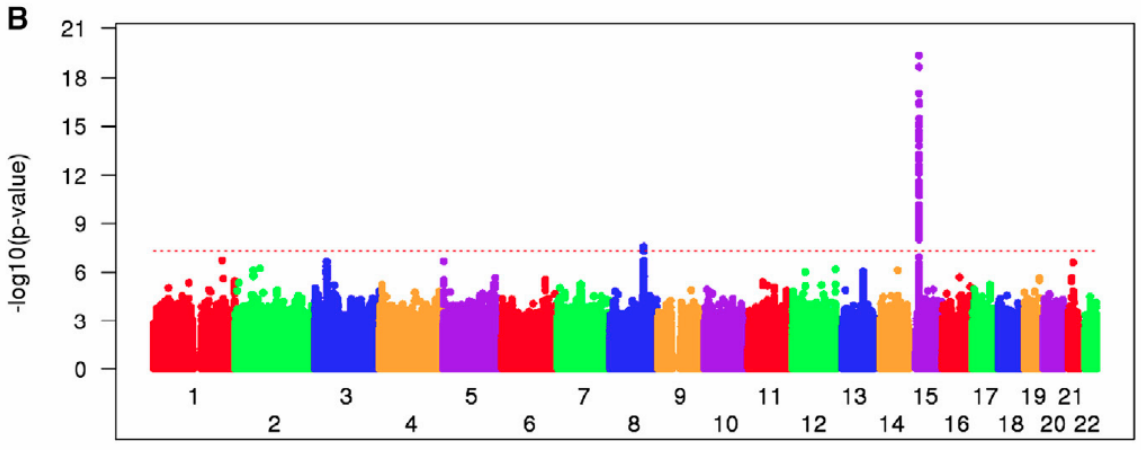


Figure 1.3. Manhattan plot of meta-analysis from GeM GWAS of HD motor onset.
Genome-wide significant peaks are seen on chromosome 15 and 8, and near-significant on chromosome 3. Reproduced from GeM-HD (2015).

The chromosome 15 signals were in or near FANCD2/FANCI-Associated Nuclease 1 (*FAN1*), a DNA endo/exonuclease involved in DNA repair. *FAN1* is highly expressed in the brain (GTEx, 2015) and the minor allele at rs2140734 is significantly associated with reduced expression in liver ($p = 5.2 \times 10^{-7}$), a tissue that also demonstrates somatic instability (Wheeler, 1999, Mangiarini et al., 1997). At the chromosome 8 locus, the two main candidates are *RRM2B* and *UBR5*. RRM2B is an enzyme involved in dNTP synthesis (Pontarin et al., 2011), which is important during DNA replication and repair (Pontarin et al., 2012), as well as regulating mitochondrial DNA content (Bourdon et al., 2007) and suppressing the oxidative stress pathway (Kuo et al., 2012). UBR5 is a ubiquitin ligase that tags proteins for proteasomal degradation, and has been investigated for its role in polyglutamine protein aggregation (Ortega and Lucas, 2014). The chromosome 3 signal is likely underlain by *MLH1* or *LRRFIP2*. *MLH1* which is known to modify somatic instability in HD mice (Pinto et al., 2013a) and directly interacts with FAN1, lending further support to the chromosome 15 peak. Dominant loss of function mutations in *MLH1* are associated with hereditary bowel cancer, in which tumours display instability of dinucleotide repeats (Pal et al., 2008). LRRFIP2 is involved in protein-protein interactions that regulate Wnt-signalling (Liu et al., 2005).

DNA repair pathways were also associated with motor onset in the GeM gene set enrichment analysis (GSEA). The most significant repair pathway, GO:33683 *nucleotide-excision repair, DNA incision* ($p = 1.69 \times 10^{-6}$), contains *FAN1*. It also remained significant when *FAN1* was excluded, thereby implicating the wider suite of DNA repair proteins in HD pathogenesis.

Lee et al. (2017) extended the GeM-HD GWAS by genotyping a further 3,314 subjects, confirming the chromosome 15 and 8 signals, and raising the chromosome 3 locus to genome-wide significance, suggesting *MLH1* variation may delay onset by reducing expression. Bettencourt et al. (2016) showed that variants in DNA repair genes from the GeM-HD GWAS, including *FAN1* and *RRM2B*, also influenced onset in the other polyglutamine diseases, suggesting a common mechanism operates in diseases caused by CAG repeat expansion (see Chapter 3).

Hensman Moss et al. (2017b) derived a novel progression score based on principal component analysis of longitudinal motor, cognitive and imaging measures in 218 HD subjects from Track-HD. They conducted a GWAS in the Track-HD (n = 216) and Registry (n = 1773) cohorts, identifying a chromosome 5 locus that slowed disease progression. The signal spans three genes; *MSH3*, *DHFR* and *MTRNR2L2*. rs557874766, the lead SNP in TRACK-HD, which was imputed, was genome-

wide significant in the meta-analysis (p=1.58E-08) and encodes the amino acid change p.P67A in *MSH3* (see Chapter 8). Colocalisation analyses with GTEx data suggested the SNPs driving this signal were eQTLs for both *MSH3* and *DHFR*, reducing expression in brain and peripheral tissues. They conclude that the modifier effect could be due to functional or expression change in *MSH3*, *DHFR* or both. It is notable that MSH3 promotes somatic expansion in animal models of HD (Tome et al., 2013a, Dragileva et al., 2009, Williams and Surtees, 2015), DM1 (Nakatani et al., 2015b, Stevens et al., 2013, Du et al., 2013b, Seriola et al., 2011b, Nakatani et al., 2015c, Williams and Surtees, 2015, Kantartzis et al., 2012, Dragileva et al., 2009, Tome et al., 2013a, van den Broek et al., 2002, Foiry et al., 2006) and Friedreich's ataxia (Bourn et al., 2012, Zhao et al., 2015b, Ezzatizadeh et al., 2012), as well as in DM1 patients (Morales et al., 2016), whereas DHFR does not. The study also identified the chromosome 15 signal at *FAN1* and *MTMR10*, and the chromosome 3 signal at *MLH1*, which were just below the threshold of genome-wide significance. Their pathway analysis highlighted DNA repair, particularly mismatch repair.

Genetic modifiers remain a priority in order to improve our understanding of pathogenesis, to enable clinical trials that stratify patients accounting for natural genetic variability, and to provide novel therapeutic targets. Though the natural modifying effect of a variant may be small, pharmacological manipulation of the pathway could result in a stronger effect.

1.3 DNA repair

1.3.1 The DNA damage response

DNA is continually damaged, for example by UV sunlight, ionising radiation, chemical mutagens and, importantly in the central nervous system, by endogenous metabolic processes producing reactive oxygen species, all of which could result in cancer or cell death if unrepaired. The DNA damage response (DDR) is a series of overlapping signalling pathways that sense and set repair in motion. Mispaired bases are replaced with correct ones by mismatch repair (MMR), chemically altered bases are excised and replaced by base excision repair (BER), and more complex lesions like pyrimidine dimers are corrected by the removal of an oligonucleotide through nucleotide excision repair (NER) which is versatile because it senses structural distortions in DNA rather than specific base modifications. Covalent links between bases on different DNA strands are repaired by interstrand crosslink (ICL) repair, breaks in one DNA strand are repaired by single-strand break repair (SSBR), and double-strand breaks (DSB) are processed by either homologous recombination (HR), which during DNA replication uses the sister chromatid as a template, is precise and occurs in mitotic neural progenitor cells, or nonhomologous end joining (NHEJ), in which broken DNA ends are directly ligated, is error-prone and predominates in post-mitotic neurons (Ciccia and Elledge, 2010). Sensor proteins recognise specific DNA lesions, then the signal is amplified by phosphorylation of phosphatidylinositol 3-kinase-like protein kinases (PIKKs) such as ATM, ATR and DNA-PK, or poly(ADP-ribose) polymerases (PARPs) to recruit effector proteins through a cascade of various posttranslational modifications. Interestingly, even in the absence of DNA damage, tethering of sensor proteins to chromatin or compaction of chromatin itself are sufficient to activate the DDR (Ciccia and Elledge, 2010, Burgess et al., 2014).

1.3.1.1 Mismatch repair

1.3.1.1.1 Function

During DNA replication, at microsatellites the template and daughter strands can dissociate and reanneal incorrectly, causing the number of repeat units on each strand to differ, with the unpaired nucleotide partially extrahelical in what is

known as an insertion-deletion loop (IDL). Together with base mismatches, which are caused by DNA polymerase errors that escape proofreading, these are repaired by the mismatch repair (MMR) system, which degrades the erroneous strand, allowing DNA polymerase to resynthesise an error free copy of the template. In the absence of MMR, IDLs and mismatches go uncorrected, leading to microsatellite instability (MSI) and eventually cancer (Jiricny, 2006).

1.3.1.1.2 Bacterial mismatch repair

Mismatch repair (MMR) proteins were named 'Mut' because inactivation in *E. coli* resulted in hypermutable strains. In *E. coli* the mismatch is recognised by MutS, which then forms a complex with MutL and MutH. The MutH endonuclease cleaves the newly synthesised unmethylated strand. MutL and MutS then act with an exonuclease and helicase to excise the DNA between the break and the mismatch, and the resulting gap is filled by DNA polymerase and ligase (Cooper, 2000).

1.3.1.1.3 Eukaryotic mismatch repair

Eukaryotes, including humans, have a similar MMR system. Mammalian cells recognise the newly synthesised strand by the presence of single strand breaks (SSB) or insertion-deletion loops (IDL). Humans have two MutS homologues, MutSα (MSH2/MSH6) which targets a base mispair or 1-2 unpaired bases, and MutSβ (MSH2/MSH3) which targets small insertion-deletion loops (IDL) of 1-15 nucleotides, as well as DNA with a 3´ single-stranded overhang (Pearl et al., 2015, Iyer et al., 2015, Gupta et al., 2011b). ATP binding induces a conformational change which allows MutS to move along the DNA as a sliding clamp (Gradia et al., 1997). Three MutL homologues perform the function of prokaryotic MutL and MutH, cleaving the DNA of the lesioned strand, namely MutLα (MLH1/PMS2), MutLβ (MLH1/PMS1) and MutLγ (MLH1/MLH3). MutLα is the most active and endonucleolytically cleaves the lesioned strand near the mismatch in a PCNA, RFC and ATP-dependent process (Muro et al., 2015, Xiao et al., 2014). The MutS-MutL complex recruits PCNA and the endonuclease EXO1 to excise the cleaved strand, which is then resynthesised by DNA polymerase δ (*POLD*) and the repair process is completed by DNA ligase 1 (LIG1) (Muro et al., 2015).

1.3.1.1.4 Disease

Inactivation of MutS and L homologues by mutation or reduced expression of MutLα by promoter hypermethylation result in hereditary nonpolyposis colorectal cancer (HNPCC or Lynch syndrome), a relatively common inherited cancer syndrome (OMIM, 2015) causing tumours of colorectal, endometrial, ovarian, stomach, small bowel, hepatobiliary, urinary tract and skin tissue. Around 70-90% of cases are caused by mutations in *MLH1* and *MSH2*, with *MSH6* and *PMS2* mutations accounting for the remainder (Muro et al., 2015). MutSβ causes repeat instability, but evidence for MutSα is less consistent (Iyer et al., 2015). The phenotype of $Msh6^{-/-}$ mice is less severe compared to $Msh2^{-/-}$ animals because MutSβ can deal with most IDLs (de Wind et al., 1999). *MSH3* mutation itself has not been linked to cancer in humans, but loss of *MSH3* in tumour cells is correlated with increased microsatellite instability (Haugen et al., 2008), *Msh3* knockout mice develop cancers only late in life, and *Msh3*/*Msh6* double knockout increases cancer susceptibility more than single knockout of either gene (de Wind et al., 1999, Edelmann et al., 2000). $Msh3^{-/-}$ animals are not tumour prone, most likely because MutSα can initiate repair of most replication errors (Edelmann et al., 2000, Jiricny, 2006).

1.3.1.1.5 Mismatch repair assays

MMR deficient cells are resistant to methylating agents that generate O^6-methylguanine (^{Me}G), which pairs with C or T during replication. These mispairs are recognised by MMR machinery, but as the modified base is on the template strand,

MMR removes the normal base and the repair polymerase regenerates the mispair, repeatedly triggering MMR until the replication fork arrests. MMR-deficient cells do not attempt to process these mispairs, so survive at the cost of extensive mutagenesis (Jiricny, 2006). MMR-deficient cells are also tolerant to 6-thioguanine (6-TG), which is incorporated into DNA, methylated to 6-methythioguanine (6-^{Me}TG), and acts in a similar way to ^{Me}G, being recognised predominantly by MutSα. Interestingly, MMR-deficient cells may also be more sensitive to ICL-induced cell death, suggesting functional interplay between these two repair pathways (Jiricny, 2006, Fiumicino et al., 2000).

1.3.1.2 Base excision repair

Bases damaged by oxidation, deamination or alkylation can cause mispairing and mutation. They are recognised and removed by DNA glycosylases, such as OGG1 which identifies 8-oxoguanine, forming apurinic/apyrimidinic (AP) sites. These are then cleaved by an AP endonuclease, resulting in a single-strand break that is processed by either short-patch repair, to replace a single nucleotide, or long-patch repair, to synthesise up to 10 nucleotides. *Ogg1* deficient HD mice show reduced somatic expansion and delayed symptom onset, and treatment with a reactive oxygen species scavenger improves motor phenotype (Budworth et al., 2015). Fen1, which removes the 5' flap generated during long patch BER, may also be involved in the generation of repeat expansions (Liu and Wilson, 2012).

1.3.1.3 Interstrand crosslink repair

ICLs are extremely toxic because they block progression of replication and transcription machinery, resulting in replication fork collapse, double strand breaks and chromosomal destabilisation (Noll et al., 2006, McCabe et al., 2009). Three models of ICL repair have been proposed (Raschle et al., 2008, Wang, 2007, Huang et al., 2013).

1. In **replication-coupled ICL repair**, which involves the FA pathway, replication forks collide, either on one or both sides of an ICL. The FANCM-FAAP24-MHF complex recognises the stall and recruits the FA core complex of 8 proteins (FANCA, B, C, E, F, G, L, and M). FANCM activates the ATR kinase, which phosphorylates the FA core complex and ID2 complex (FANCD2-FANCI). The FA core ubiquitinates the ID2 complex, which recruits structure-specific nucleases, most likely SLX1, SLX4, XPF and ERCC1, to make single strand incisions either side to unhook the crosslink. Translesion synthesis (TLS) DNA polymerases can then bypass the lesion and the replication fork is restored by strand invasion and homologous recombination (Jin and Cho, 2017, Kee and D'Andrea, 2010, Kim and D'Andrea, 2012, Zhang and Walter, 2014, Huang and D'Andrea, 2010). The cross link is then removed by nucleotide excision repair (NER).
2. In **repair-independent replication**, replication forks bypass the ICL without repair. The stalled fork is recognised by the FANCM/MHF complex translocase, which moves it past the lesion (Meetei et al., 2005, Huang et al., 2013).
3. In the **FA-independent pathway**, the DNA glycosylase NEIL3 cleaves the N-glycosidic bond between bases and the sugar-phosphate backbone, allowing unhooking, then gap filling by TLS polymerases. This avoids double strand breaks, minimising the chance of chromosomal rearrangements (Rolseth et al., 2013, Wang et al., 2016).

Fanconi anaemia is caused by a mutation in one of the 17 known FANC genes, leading to failure of ICL repair (Ceccaldi et al., 2016a). Cells are susceptible to ICL-inducing agents and most patients develop cancer, commonly acute myeloid leukaemia, as well as bone marrow failure, congenital defects, skin pigmentation and endocrine abnormalities. Inheritance is usually autosomal recessive. Treatment with androgens and haematopoietic growth factors can transiently help bone marrow failure, though long-term treatment involves bone marrow transplantation.

The role of the FA pathway in repeat instability has not yet been explored, though recently FAN1 was shown to protect against expansion in a fragile X mouse model (Zhao and Usdin, 2018).

1.3.2 DNA repair and neurodegenerative disease

1.3.2.1 Susceptibility of nervous system tissue to DNA damage

DNA repair defects have long been linked to cancer syndromes, but many, for example those caused by mutations in components of NER like xeroderma pigmentosum, double strand break repair (DSBR) such as ataxia telangiectasia, or SSBR like ataxia with oculomotor apraxia-1 (AOA1), display neurodegeneration, including microcephaly, cognitive impairment, deafness, ataxia and neuropathy. This suggests the nervous system is especially sensitive to DNA damage, though the reason for selective vulnerability of postmitotic neurons remains unclear.

1.3.2.2 Oxidative stress

The nervous system is highly dependent on oxidative metabolism, which generates free radicals that can cause DNA strand breaks (McKinnon, 2009) and can promote repeat expansion (Kovtun et al., 2007, McMurray, 2008). The brain metabolises around 20% of consumed oxygen, but has a lower capacity than other tissues to neutralise reactive oxygen species and neurons are particularly vulnerable to oxidative stress (McKinnon, 2009, Canugovi et al., 2013). Increased levels of DNA damage such as strand breaks and oxidative lesions have been reported in human Alzheimer's, Parkinson's and amyotrophic lateral sclerosis (ALS) brain, and lower levels of DDR proteins are seen in AD, though it is unclear whether these are the cause or consequence of the primary neurodegenerative process. Oxidative lesions are primarily repaired by BER, and levels of glycosylases UDG1 and βOGG1, which recognise oxidised bases, are reduced in human AD brain (Canugovi et al., 2013). Genomic instability progressively increases with age due to decreasing DNA repair activity, the accumulation of irreversible mutations through erroneous repair of DNA lesions, and the failure of chromatin to return to its predamaged conformation, all of which could contribute to age-related neurodegeneration (Madabhushi et al., 2014).

1.3.2.3 Double strand breaks

The role of DNA damage in the nervous system has long been studied in ataxia telangiectasia (AT), which is caused by mutations in *ATM*, a serine/threonine kinase recruited to double strand breaks (DSB) to coordinate repair. Compared to other lesions, DSBs are rare events, but they are extremely damaging because they can cause large chromosomal rearrangements leading to cell death or tumorigenesis (Jackson, 2002). AT is a multisystem disease with radiosensitivity, immunodeficiency and cancer, but it also causes cerebellar degeneration. The rare A-T like disease (ATLD) is a very similar syndrome caused by mutations in *MRE11*, which is involved in DSB repair. The similarities suggest defective DSB repair causes neurodegeneration, though the reason for selective vulnerability remains unclear (Madabhushi et al., 2014). DSBs are produced in neurons during normal physiological activity, though it is unclear whether these serve a purpose in learning new tasks or are the result of neuronal activity (Suberbielle et al., 2013).

1.3.2.4 Single strand breaks

Single strand breaks (SSB) are three times commoner than DSBs and can also provoke apoptosis, but whereas proliferating cells can repair these through HR during DNA replication, non-proliferating cells have fewer options available. Ataxia with oculomotor apraxia-1 (AOA1) involves cerebellar degeneration, cognitive impairment, low albumin and cholesterol and is caused by mutation in *APTX*, which processes DNA ends in SSBR and also interacts with DSB

machinery such as XRCC4 (Clements et al., 2004). Spinocerebellar ataxia with axonal neuropathy (SCAN1) is a rare disease of cerebellar degeneration and peripheral neuropathy caused by mutation in *TDP1*, a phosphodiesterase which acts on the DNA ends of SSBs and DSBs, leading to the accumulation of SSBs (El-Khamisy et al., 2005). These conditions demonstrate that failure of SSB repair can also result in neurodegeneration.

1.3.2.5 Ageing

With increasing age, DNA damage and mutations accumulate, and DNA repair activity declines (Lu et al., 2004). In mice, liver mutations almost quadruple with age (Dolle et al., 1997) and in adult human frontal cortex up to 40% of neurons have copy number variations (CNVs) (McConnell et al., 2013). Following DNA repair, chromatin may not return to its predamaged state, which may affect expression with age (Madabhushi et al., 2014).

The DDR is important both during neural development and in mature neurons. Mutations in DNA repair factors cause severe neurodevelopmental disorders as well as age-related neurodegeneration. Postmitotic neurons are particularly vulnerable and acquire DNA damage such as oxidation and strand breaks with age, potentially because of declining repair activity. One of the greatest challenges in HD will be understanding how on the one hand DNA repair guards genomic stability, whilst on the other hand contributing to cell death (Jiricny, 2006).

1.3.3 Modifiers of repeat stability

1.3.3.1 DNA repair in Huntington's disease

Several DNA maintenance processes have been implicated in causing repeat expansion, including DNA replication, base excision repair, double strand break repair, nucleotide excision repair and recombination (Castel et al., 2010, Pearson et al., 2005a), but MMR is the strongest driver (Castel et al., 2010, Slean et al., 2008). DNA repair proteins appear to act on repeat sequences once they reach a threshold length. In Huntington's disease mouse models, somatic expansion increases with age (Mangiarini et al., 1997, Wheeler, 1999, Ishiguro et al., 2001, Lee et al., 2011a), modifies disease course (Kovalenko et al., 2012, Wheeler, 2003), is exacerbated by oxidative DNA damage (Kovtun et al., 2007, Bogdanov et al., 2001), depends on a functional mismatch repair system (Manley et al., 1999, Tome et al., 2013a, Pinto et al., 2013a), and can be ameliorated by manipulating DNA repair genes (Wheeler et al., 2003, Tome et al., 2013a, Kovtun et al., 2007). Knockout of *MSH2* (Lopez Castel et al., 2010, Manley et al., 1999), *MLH1* or *MLH3* (Pinto et al., 2013a) completely ablates and knockout of *MSH3* (Dragileva et al., 2009, Tome et al., 2013a) and *PMS2* (Gomes-Pereira, 2004, Gomes-Pereira et al., 2014b, Pinto et al., 2013b) reduces repeat expansion. CAG expansions are therefore driven by the mismatch repair complexes **MutSβ** (MSH2-MSH3), **MutLα** (MLH1-PMS2) and **MutLγ** (MLH1-MLH3). The inculpation of MutSβ, which deals with short insertion-deletion loops, suggests expansion may result from short incremental expansions rather than large jumps (Schmidt and Pearson, 2016, Williams and Surtees, 2015). Slip-outs may have unpaired or mispaired nucleotides at the junction, which could confuse the MMR system, potentially targeting repair to the incorrect strand and resulting in an expansion (Schmidt and Pearson, 2016). Knockout of *OGG1*, a base excision repair protein, also reduced expansion and delayed onset in HD mice (Budworth et al., 2015).

It may not be necessary to completely inactivate MMR proteins, as subtle genetic variation also affects instability; germline and somatic instability were noted to differ between strains of HD mice (Tome et al., 2013a). This variability was mapped to variants in *Msh3* that influenced its expression level. A genome-wide association study also identified *Mlh1* as an influence on strain-specific variation in instability (Pinto et al., 2013a). Prognosis in human repeat expansion

diseases may, therefore, be tempered by the presence of variants in DNA repair genes. Elucidating a mechanism through which MMR could switch from expansion to contraction could have therapeutic potential (Lopez Castel et al., 2010).

1.3.3.2 DNA repair in other repeat expansion diseases

DNA repair is known to modify repeat instability in numerous repeat expansion diseases. In myotonic dystrophy, the MMR complex MutSβ (Msh2/Msh3) is required for expansion (Nakatani et al., 2015b, Stevens et al., 2013, Du et al., 2013b, Seriola et al., 2011b, Nakatani et al., 2015c, Williams and Surtees, 2015, Kantartzis et al., 2012, Dragileva et al., 2009, Tome et al., 2013a, van den Broek et al., 2002, Foiry et al., 2006) and *MSH3* polymorphisms are associated with variation of somatic instability in patient blood (Morales et al., 2016). In SCA3, ERCC6, a base excision repair protein, is associated with intergenerational repeat expansion (Martins et al., 2014). In fragile X mice, Msh2 is required for expansion (Lokanga et al., 2014) and Fan1 is protective (Zhao and Usdin, 2018), and in fragile X patients, DNA repair genes are downregulated in blood (Xu et al., 2013). In Friedreich's ataxia mice, Msh2, Msh3, Msh6 and Pms2 (Bourn et al., 2012, Zhao et al., 2015b, Ezzatizadeh et al., 2012) have been associated with expansion.

1.3.3.3 Mechanisms underlying repeat instability

Repetitive DNA sequences form unusual non-canonical structures, including slipped strands, hairpin loops, G-quadruplexes and R-loops (Mirkin, 2007, Neil et al., 2017, McMurray, 2010), the stability of which correlates with expansion (Gacy et al., 1995). Hairpins contain an A:A base pair mismatch in the stem that is predicted *in silico* to result in a Z DNA structure, with double helix winding to the left instead of the right, perhaps with flipping out of the mismatched bases (Khan et al., 2015). Such structures could form substrates for the DNA mismatch repair machinery. In DM1 patients, for example, higher levels of slipped strand DNA are found in tissues with the most repeat instability, such as the heart (Axford et al., 2013). The disease process itself may also induce DNA damage, either through a direct effect of the expanded repeat or through mitochondrial dysfunction and excitotoxicity, which activate the DNA damage response (Shah and Mirkin, 2015). Processes involved in germline and somatic and instability may differ, with the former more related to DNA replication during cell division and the latter, which occurs in non-dividing neurons, associated with DNA repair, transcription and chromatin dynamics.

1.4 FAN1

FAN1 is a highly conserved DNA endo- and exonuclease originally described in 2010 by four groups (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Smogorzewska et al., 2010a). It is required for the Fanconi anaemia (FA) interstrand crosslink repair (ICL) pathway , acting in complex with some mismatch repair (MMR) proteins (MacKay et al., 2010b, Kratz et al., 2010a, Liu et al., 2010b, Smogorzewska et al., 2010a), though its precise role in this process remains unclear (Thongthip et al., 2016, Lachaud et al., 2016a, Lachaud et al., 2016b). FAN1 is structure rather than sequence specific, binding and cleaving branched 5' flap structures that occur as DNA repair intermediates (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Pennell et al., 2014, Liu et al., 2010c, MacKay et al., 2010a). It also has an independent role maintaining genomic stability and preventing chromosomal abnormalities, possibly through the regulation of replication fork dynamics (Lachaud et al., 2016a, Chaudhury et al., 2014).

1.4.1 Structure

Four domains have been characterised. Through its N-terminal ubiquitin binding domain (UBZ), monoubiquitinated FANCD2 and FANCI of the FA pathway recruit it to nuclear ICL damage foci (Liu et al., 2010c, Smogorzewska et al., 2010a), hence its name (FANCD2 and FANCI Associated Nuclease 1). Its DNA binding (SAP) domain may be involved in recruiting FAN1 to ICL damage foci (Thongthip et al., 2016). The tetratricopeptide repeat (TPR) mediates protein-protein interactions and the assembly of multiprotein complexes. Finally, its nuclease domain, a viral replication and repair nuclease (VRR Nuc), has endonuclease activity at 5' flap structures and 5'-3' exonuclease activity (MacKay et al., 2010b). FAN1's crystal structure has been determined bound to DNA substrates and suggests it may form a dimer to orient and nick DNA (Wang et al., 2014b, Zhao et al., 2014, Gwon et al., 2014, Yan et al., 2015). Interestingly, the bacterial and unicellular eukaryotic FAN1 lacks the UBZ domain (Jin and Cho, 2017).

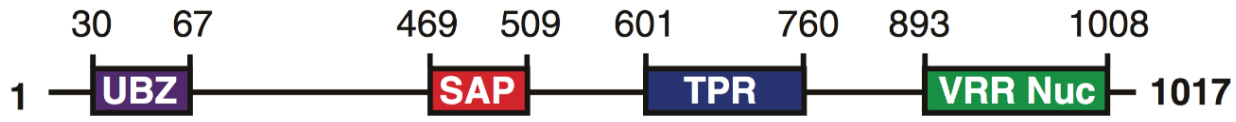


Figure 1.4. Schematic representation of FAN1.
The UBZ (ubiquitin-binding zinc finger 4), SAP (SAF-A/B, Acinus and PIAS), TRP (tetratricopeptide repeat) and VRR Nuc (viral replication and repair nuclease) motifs of hFAN1 are indicated. Reproduced from Takahashi et al. (2015)

A number of FAN1-DNA complex structures have been reported, all lacking the UBZ domain and binding DNA either as a monomer or dimer, but despite these, the mechanism of FAN1-mediated ICL repair remains unclear (Gwon et al., 2014, Wang et al., 2014a, Zhao et al., 2014, Shereda et al., 2010). FAN1 forms a bi-lobed structure with the N and C-terminal domains (NTD, CTD) positioned orthogonally. The NTD recognises the prenick duplex and the CTD interacts with the postnick duplex, but all domains are involved in binding DNA. The nuclease makes cuts at every third nucleotide of the flap, generating a 12 nt gap containing ssDNA, though it also has the ability to cleave between each third nucleotide too. FAN1 can assemble as a head-to-tail dimer in the presence of DNA through interaction between TPR and VRR domains of one molecule binding the SAP domain of another. The first molecule has the primary cleavage function, with the second involved in orientating the substrate. Monomeric and dimeric FAN1 can both cleave short DNA flaps, but the dimeric form is optimal for cleaving longer flaps (Rao et al., 2018). However, relative functions of the monomer and dimer are unknown (Zhao et al., 2014).

1.4.2 Function

1.4.2.1 Interstrand crosslink repair

FAN1 is recruited to ICLs biphasically, with an initial rapid rise through the SAP domain directly binding DNA, followed by a steady build up mediated by the UBZ domain's interaction with monoubiquitinated FANCD2 (Thongthip et al., 2016, Smogorzewska et al., 2010a, MacKay et al., 2010b). ICL repair involves the formation of double strand breaks and their repair by homologous recombination (HR) (Hanada et al., 2006, Hanada et al., 2007). Single stranded DNA is coated by replication protein A (RPA), which can be visualised as nuclear foci. During HR, RAD51 displaces RPA to generate strands that invade the sister chromatid (West, 2003). FAN1 depletion does not prevent the formation of RPA nuclear foci, but delays the disappearance of RAD51 suggesting a role in the completion of HR (MacKay et al., 2010a, Kratz et al., 2010b). γ-H2AX binds double strand breaks (DSBs) to recruit DNA repair proteins (Niedernhofer et al., 2004, Rothfuss and

Grompe, 2004). FAN1 knockdown does not impair the formation of γ-H2AX foci, but does delay their resolution (MacKay et al., 2010a, Kratz et al., 2010b), consistent with a role in resolving DSBs induced during ICL repair. These observations suggest that FAN1 participates in HR and its absence results in erroneous processing of ICLs, resulting in chromosomal aberrations (Kratz et al., 2010b, Raschle et al., 2008).

FAN1 recognises and binds the ss/dsDNA junction of 5' flap structures 3-4 nucleotides into the double stranded portion (Takahashi et al., 2015, Pizzolato et al., 2015), a region not covered by the RPA that stabilises ssDNA. It then makes an incision 2-4 nt into the dsDNA, successively cutting every third nucleotide (Wang et al., 2014b, Wang et al., 2014a), which may allow it to traverse an ICL in order to unhook it (Pizzolato et al., 2015). It has diverse nuclease activity, which would allow it to participate in several DNA repair processes, including ICL unhooking, trimming of unhooked strands and D-loop incision during HR (Jin and Cho, 2017).

However, FAN1 appears to act independently of the FA pathway, and may have additional functions in maintaining genomic stability. Lachaud et al. (2016a) showed that p.C44A/C47A UBZ domain mutant FAN1, which is unable to interact with ubiquitinated FANCD2, fully rescues ICL repair in FAN1 knockout cells, suggesting FA pathway-mediated FAN1 recruitment is dispensable for ICL repair. FAN1 can be directly recruited to ICLs via its SAP DNA binding domain (Thongthip et al., 2016) and can also act alone to efficiently unhook ICLs *in vitro* (Wang et al., 2014a, Pizzolato et al., 2015).

1.4.2.2 Replication fork recovery

Though FAN1 is able to repair ICLs independent of the FA pathway, it appears that its interaction with FANCD2 is required for genomic stability (Chen et al., 2015, Schlacher et al., 2012). The pancreatic cancer-causing UBZ domain mutation p.M50R (Smith et al., 2016) impairs recruitment by FANCD2 and has normal ICL repair function, but cells develop chromosomal abnormalities (Lachaud et al., 2016a). FAN1 nuclease and UBZ domain mutants cannot protectively restrain stalled replication forks (Ray Chaudhuri et al., 2012, Ge and Blow, 2010, Lachaud et al., 2016a), potentially leading to replication fork reversal that has previously been implicated in repeat instability (Follonier et al., 2013, McMurray, 2010, Mirkin, 2007, Ray Chaudhuri et al., 2012). Replication forks can be stalled by ICLs, a lack of nucleotides, polymerase inhibition, abasic sites, modified bases, strand breaks or abnormal secondary structures on the template strand, such as G-quadruplexes (Porro et al., 2017).

FANCD2 regulates FAN1 activity at stalled forks, protecting nascent DNA from nucleolytic degradation (Chaudhury et al., 2014). In the presence of FANCD2, FAN1 is recruited to stalled replication forks and acts in concert with MRE11 and BLM to suppress firing of new origins and promote replication fork restart. In the absence of FANCD2, FAN1 is still recruited to stalled forks, but restart does not occur efficiently, new origins are triggered and uncontrolled FAN1 degrades nascent DNA strands behind the fork. Inappropriate incision such as this could cause genomic instability rather than protect against it. Therefore, FAN1 joins a group of fork restart proteins, including FANCD2, BRCA1, MRE11, XRCC3, RAD51, CTIP, and MUS81 (Chaudhury et al., 2014), all of which have been implicated in repair of double strand breaks by homologous recombination. The main role of FANCD2 may be to protect nascent DNA strands at replication forks and ICLs from nucleolytic degradation by loading them with RAD51 (Chaudhury et al., 2014). As a nuclease that can cleave several different DNA structures, FAN1 activity likely needs to be closely controlled (Porro et al., 2017). It seems that FAN1 is important for controlling the restart of replication forks stalled not only by ICLs, but by other lesions as well (Porro et al., 2017, Zhao et al., 2014).

Porro et al. (2017) identified a PCNA interaction motif (PIP), which together with the UBZ domain recruits FAN1 to ubiquitylated PCNA at stalled replication forks, thereby preventing collapse and regulating fork progression. This PIP domain is not required for FAN1 recruitment to MMC-induced ICLs.

1.4.2.3 Protein interactions

FAN1 is known to directly interact with the ID complex (FANCD2 and FANCI), through which it is involved in ICL repair (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Smogorzewska et al., 2010a). However, it also interacts with mismatch repair complexes MutLα (MLH1 and PMS2) and MutLγ (MLH1/MLH3), which are required for repeat expansion, though the function of this interaction remains unknown (MacKay et al., 2010a, Kratz et al., 2010a, Liu et al., 2010c, Smogorzewska et al., 2010b).

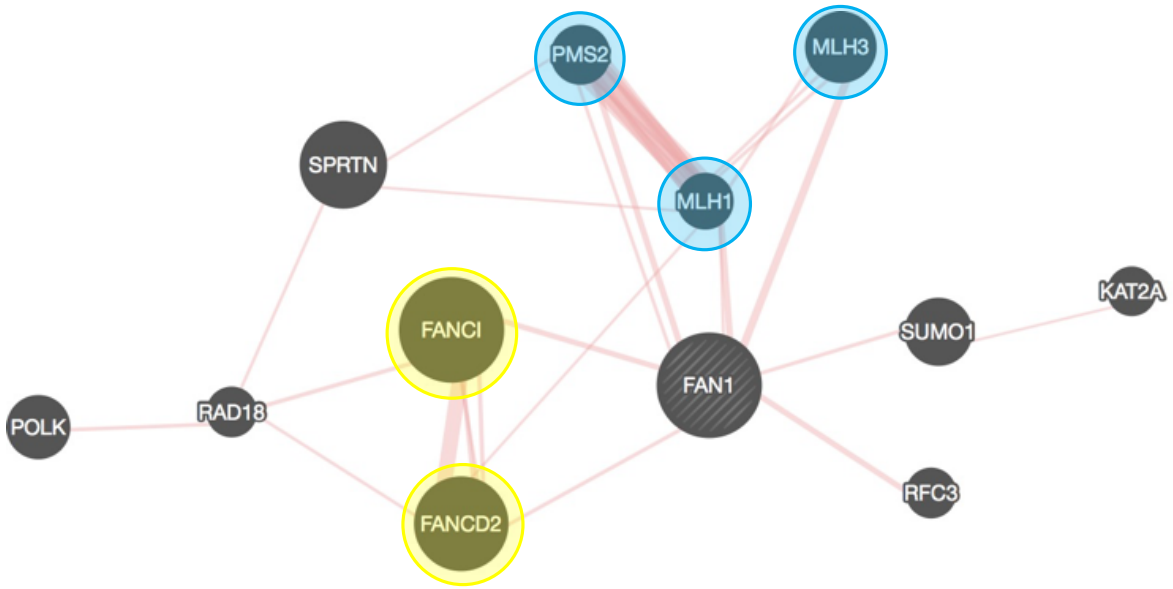


Figure 1.5. FAN1 interactome.
Blue represents mismatch repair components, yellow is Fanconi anaemia pathway components, lines represent physical interactions. Prepared using GeneMANIA (Warde-Farley et al., 2010).

1.4.2.4 Summary

Taken together, a model is emerging in which abnormal DNA structures formed by the *HTT* CAG repeat could stall replication forks. Stalled forks are then bound by PCNA, FAN1 is recruited and its nuclease activity is involved in correct restart of the fork, with FANCD2 protecting nascent DNA from degradation.

1.4.3 FAN1 depletion

Loss of FAN1 sensitises cells to ICL-inducing agents, such as mitomycin C (MMC) and cisplatin, and results in chromosomal breaks reminiscent of FA patients (Kratz et al., 2010b, Liu et al., 2010c, MacKay et al., 2010b, Akkari et al., 2000, MacKay et al., 2010a). Zhao and Usdin (2018) recently showed that FAN1 knockout in a fragile X mouse model accelerated CGG repeat expansion, particularly in liver and brain

FAN1 mutations do not cause Fanconi anaemia, but homozygous loss of function mutations result in the recessive renal disease karyomegalic interstitial nephritis (KIN), characterised by renal fibrosis, tubular degeneration and polyploidy in multiple tissues (Zhou et al., 2012, Lachaud et al., 2016b, Thongthip et al., 2016). Heterozygous truncating mutations

have been linked to pancreatic (Smith et al., 2016) and hereditary colorectal cancers (Segui et al., 2015b). *FAN1* lies in a 2 Mb region of copy number variation (CNV) due to non-allelic homologous recombination of flanking repeats. Deletion and duplication of the region have been associated with intellectual disability, epilepsy, autism and schizophrenia (Ionita-Laza et al., 2014). It is likely that chromosomal abnormalities due to failure of replication fork protection, as discussed above, underlie some of these conditions.

Disease causing mutations near the active site, such as p.D960A, abolish nuclease activity and have been found in patients with KIN, pancreatic or colorectal cancer, and schizophrenia or autism (Zhao et al., 2014). Mutations on the surface of the protein, such as p.R507H and p.P894S, may affect protein-protein interactions. Mutations elsewhere may affect the protein structure or its interaction with DNA, such as p.R377W at the interface of the helical and CTD. Generally, the KIN-causing mutations are clustered in the nuclease-containing CTD and those associated with cancer either affect the UBZ or nuclease domains (Zhao et al., 2014).

1.5 MSH3

1.5.1 Function

MSH3 forms the neuronally expressed heterodimeric complex MutSβ with MSH2, acting in the DNA mismatch repair (MMR) pathway, recognising base mismatches and small insert-deletion loops (IDL) (Tome et al., 2013a, Gonitel et al., 2008). DNA repair pathways are highly interconnected and MutSβ is implicated in at least three processes (Ashburner et al., 2000); repair of insertion-deletion loops, repair of double strand breaks by binding 3' overhangs, and repair of single-strand annealing (Lyndaker and Alani, 2009, Schmidt and Pearson, 2016). The MSH proteins are ATPases that possess the Walker ATP-binding motif, which is highly conserved in DNA repair proteins. ATP is not required for the initial recognition of mismatches (Jiricny, 2006), but it is subsequently required for the conformational change of the MutS complex, which allows it to release the mismatch and move along the DNA in the form of a sliding clamp. In the presence of a mismatch, the MutS heterodimer clamps around the DNA like a pair of praying hands (Jiricny, 2000).

1.5.2 Structure

1.5.2.1 Domains

Eukaryotic MSH proteins have five domains homologous to the bacterial form, though unlike MSH2 and bacterial MutS, MSH3 and MSH6 also have a conserved 100-600 residue N-terminal region containing a motif for interaction with PCNA (Kunkel and Erie, 2005, Clark et al., 2007a). The MSH3 protein resembles a figure of eight, with the lower of the two channels penetrating the protein forming the DNA binding domain, and the dimerization interface and nucleotide binding domain at the opposite end of the molecule (Schofield and Hsieh, 2003). Domains I and IV bind DNA, with I directly contacting the mismatched base and IV forming jaws that clamp the protein to DNA (Schofield and Hsieh, 2003). Domain V, at the C-terminus, contains the dimerisation interface and the nucleotide binding site for ATP. Domain III provides a structural bridge between the ATPase domain and the DNA binding site. The crystal structure of MutSβ in complex with DNA insertion-deletion loops shows it binds both DNA strands 5' of the lesion through domain I, but only the loop-containing strand 3' of the lesion through domain IV (Gupta et al., 2011a). DNA binding is mediated by residues 245-246, and mutation of the homologous yeast residues leads to microsatellite instability (Schmutte et al., 2001). Tyr245 interacts with the normal strand 5' in the double stranded region, whereas Lys246 contracts the loop strand.

1.5.2.2 Protein interactions

MSH3 forms the heterodimeric MutSβ MMR complex by binding MSH2 through N-terminal residues 126-250 and C-terminal residues 1050-1128 (Acharya et al., 1996). It also interacts with EXO1, an MMR exonuclease that excises mismatch-containing DNA tracts, through residues 75-297 (Schmutte et al., 2001). MSH3 contains an N-terminal PCNA interaction motif, Qxx(LI)xxFF (PIP box, which in MSH3 is encoded by QAVLSRFF at residues p.21-28) (Kleczkowska et al., 2001, Clark et al., 2000, Flores-Rozas et al., 2000). PCNA is known to be involved in delivering MSH proteins to mismatches (Lau and Kolodner, 2003) and increases mismatch binding specificity (Flores-Rozas et al., 2000). It is a processivity factor for DNA polymerase, acting as a sliding clamp and participating in both DNA replication and MMR (Clark et al., 2000, Flores-Rozas et al., 2000, Kleczkowska et al., 2001, Goellner et al., 2015). The PCNA motif tends to be followed by a non-conserved sequence containing basic amino acids and often prolines, which interacts with the interdomain connector loop (IDCL) of PCNA (Gulbis et al., 1996), the domain through which PCNA interacts with proteins (Kleczkowska et al., 2001). PCNA-binding proteins can be divided into two groups based on the amino sequence in this region, one with a preponderance of basic residues and the other with proline residues (Zhang et al., 1999). MSH6 belongs to the former and MSH3 to the latter (Kleczkowska et al., 2001). Sequence analysis suggests these regions form short, flexible connector domains. The binding of MutSα to mismatch substrates leads to dissociation from PCNA. This suggests that MutSα and/or MutSβ may also be involved in DNA replication along with PCNA, but that they are handed over to the MMR machinery when a mismatch is detected (Lau and Kolodner, 2003, Jiricny, 2006). Mutational studies suggest the N-terminal interaction with PCNA is important for MMR (Schofield and Hsieh, 2003), with variants in the MSH3 or MSH6 PCNA binding motif strongly reducing PCNA binding (Clark et al., 2000, Flores-Rozas et al., 2000, Kleczkowska et al., 2001) and increasing mutation rates, though some MMR function is retained (Clark et al., 2000).

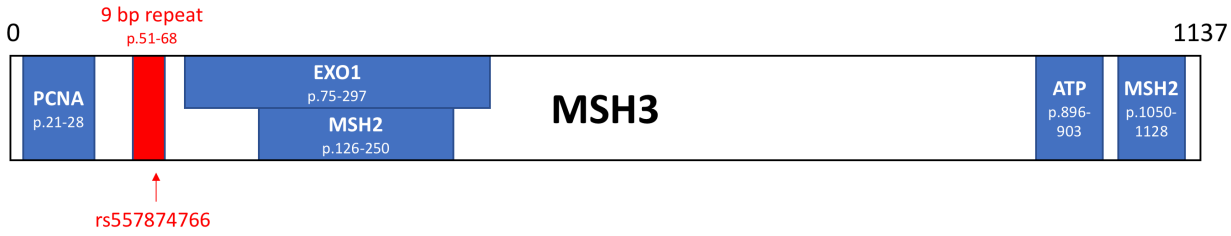


Figure 1.6. Schematic representation of MSH3.
Interaction domains are given in blue, and the 9 bp tandem repeat in red.

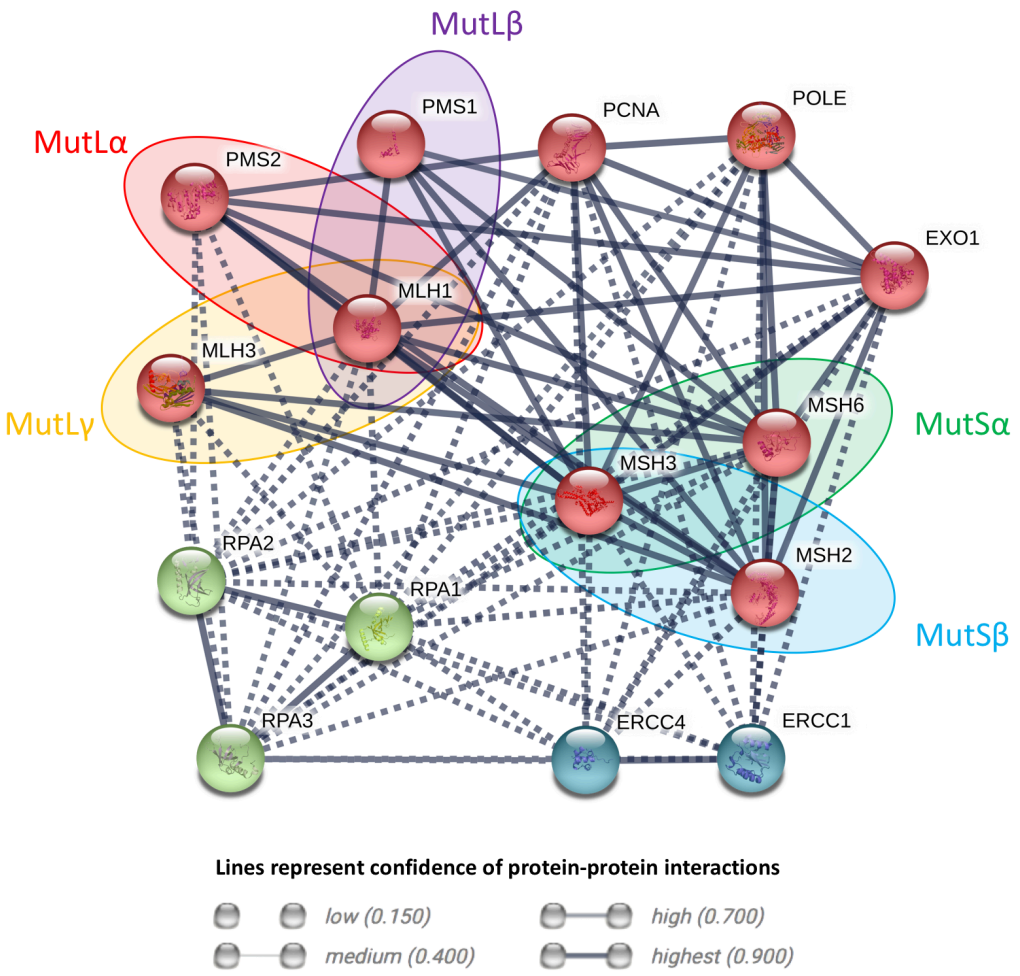


Figure 1.7. MSH3 protein interactions.

Lines represent functional interaction. Line thickness represents the confidence of the interaction. Proteins are clustered into three groups (red, green, blue) by k-means. Generated in STRING. MutS and MutL protein complexes are marked with coloured ovals.

1.5.2.3 MSH3 and DHFR share a promoter

MSH3 is situated head-to-head with *DHFR* and they share a common promoter, but are divergently transcribed (Watanabe et al., 1996, Drummond, 1999). Induction of *DHFR* expression by methotrexate, a DHFR enzyme inhibitor, leads to a co-amplification of *MSH3*, and both genes are also upregulated in childhood acute lymphoblastic leukaemia (ALL) (Watanabe et al., 1996). *DHFR* upregulation is the main cause of methotrexate resistance in ALL. Some studies have suggested that the parallel increase in MSH3 may sequester MSH2 away from MutSα, thereby impairing base mismatch repair at the expense of IDL repair (Zhang et al., 1999, Drummond, 1999, Drummond et al., 1997, Marra et al., 1998, Pandit et al., 2001, Irving and Hall, 2001, Swann et al., 1996), resulting in methotrexate-induced hypermutability, though results of DNA repair assays have conflicted (Matheson et al., 2007). It is interesting that *MSH3*, a key MMR component, shares a promoter with *DHFR*, which can influence replication fidelity through the folate-dependent biosynthesis of purines (Drummond, 1999). The promoter contains sites for Sp1 and E2F transcription factors. Expression of both *DHFR* and *MSH3* is influenced by cell cycle progression, increasing particularly during S phase with DNA replication, as well as at the onset of cell proliferation, with both genes expressed in parallel in most cases. The E2F site is close to the initiation site of *DHFR* and Sp1 is nearest the *MSH3* initiation site, but there is no direct evidence for uncoupled expression of the two genes (Drummond, 1999). However, the existence of distinct transcription factor binding sites raises the possibility

of a more complex regulator mechanism. The genomic organisation linking *DHFR* and *MSH3* may suggest communication between nucleotide synthesis and mismatch repair pathways. For example, variation in dNTP concentrations is known to influence the rate of mutation (Bebenek and Kunkel, 1990, Bebenek et al., 1992). The requirement for nucleotides rises during DNA replication, which is when mismatches are introduced and repaired, so the shared promoter arrangement may have arisen as an anticipatory mechanism to protect genome stability.

Unlike MSH3, DHFR has not been implicated in HD pathogenesis. In the R6/1 transgenic mouse model of HD, coding variation in *MSH3* that lowers its expression, as measured by western blot, reduces CAG repeat expansion (Tome et al., 2013a). Tome et al. (2013a) identified 7 polymorphisms that resulted in non-synonymous amino acid changes in exons 2, 3, 7, 8 and 10. Most residues were conserved, but only p.T321I was predicted damaging. This variant is not in a protein interaction domain, but may affect protein stability or conformation. One variant, p.A82S is within the Exo1 binding domain and is 54 amino acids downstream of the PCNA motif; the homologous human residue is p.S116. The mouse and human protein sequences are aligned below, and show 80% sequence identity (Blastp and Clustal2.1), differing mostly at the N-terminus. The authors postulated that *MSH3* transcription was unaffected because its shared promoter and *DHFR* expression were undisrupted, and that instead the low MSH3 protein level may reflect its inability to form the MutSβ protein complex with MSH2. *Msh2* knockout in mice, for example, leads to undetectable levels of MSH3 (Tome et al., 2013b).

Figure 1.8. Alignment of mouse and human MSH3 protein sequences.

The PCNA interaction motif is shown in yellow (Kleczkowska et al., 2001, Clark et al., 2000, Flores-Rozas et al., 2000), the EXO1 binding domain in cyan (Schmutte et al., 2001), the MSH2 binding domain in magenta (note EXO1 and MSH2 binding domains overlap. The residues forming salt bridges between MSH3 and MSH2 are in white on blue), the key DNA binding residues are in white on black (Gupta et al., 2011a), the ATP binding site in green, and the sites of the 7 coding variants from Tome et al. (2013a) are in red. Protein binding domains are relative to the human sequence. Sequence identity is 79% (Blastp) to 80.91% (Clustal2.1).

```
mouse       MPRGKSASGGSTAAGPGPGRQTVLSRFFRSAGSLRSSASSTEPAEKVTEG----------     50
human       MSRRKPASGGLAASSSAPARQAVLSRFFQSTGSLKSTSSSTGAADQVDPGAAAAAAAAAA     60
            * * * **** :*:. .*.**:******:*:***:*::***  *::*  *

mouse       ------------------------DSRKRSLGNGGPTKKKARKVPEKEEENISVAAHHPE     86
human       AAPPAPPAPAFPPQLPPHIATEIDRRKKRPLENDGPVKKKVKKVQQKEGGSDLGMSGNSE     120
                                     :** * *.**.***.:** :**  .    : : *

mouse       AKKCLRPRIVLKSLEKLKEFCCDSALPQNRVQTEALRERLEVLPRCTDFEDITLQRAKNA     146
human       PKKCLRTRNVSKSLEKLKEFCCDSALPQSRVQTESLQERFAVLPKCTDFDDISLLHAKNA     180
             ***** * * *****************.*****:*:**: ***:****:**:* :****

mouse       VLSEDSKSQANQKDSQF------GPCPEVF--QKTSDCKPFNKRSKSVYTPLELQYLDMK     198
human       VSSEDSKRQINQKDTTLFDLSQFGSSNTSHENLQKTASKSANKRSKSIYTPLELQYIEMK     240
            * ***** * ****: :      * .   .   :.: .*  ******:********::**

mouse       QQHKDAVLCVECGYKYRFFGEDAEIAARELNIYCHLDHNFMTASIPTHRLFVHVRRLVAK     258
human       QQHKDAVLCVECGYKYRFFGEDAEIAARELNIYCHLDHNFMTASIPTHRLFVHVRRLVAK     300
            ************************************************************

mouse       GYKVGVVKQTETAALKAIGDNKSSVFSRKLTALYTKSTLIGEDVNPLIRLDDSVNIDEVM     318
human       GYKVGVVKQTETAALKAIGDNRSSLFSRKLTALYTKSTLIGEDVNPLIKLDDAVNVDEIM     360
            **********************:**:************************:***:**:**:*

mouse       TDTSTNYLLCIYEEKENIKDKKKGNLSVGIVGVQPATGEVVFDCFQDSASRLELETRISS     378
human       TDTSTSYLLCISENKENVRDKKKGNIFIGIVGVQPATGEVVFDSFQDSASRSELETRMSS     420
            *****.***** *:***::******: :***************.******* *****:**

mouse       LQPVELLLPSDLSVPTEMLIQRATNVSVRDDRIRVERMNNTYFEYSHAFQTVTEFYAREI     438
human       LQPVELLLPSALSEQTEALIHRATSVSVQDDRIRVERMDNIYFEYSHAFQAVTEFYAKDT     480
            ********** **  ** **:***.***:*********:* *********:******::

mouse       VDSQGSQSLSGVINLEKPVICALAAVIRYLKEFNLEKMLSKPESFKQLSSGMEFMRINGT     498
human       VDIKGSQIISGIVNLEKPVICSLAAIIKYLKEFNLEKMLSKPENFKQLSSKMEFMTINGT     540
            ** :*** :**::********:***:*:***************.****** **** ****

mouse       TLRNLEILQNQTDMKTKGSLLWVLDHTKTSFGRRKLKNWVTQPLLKLREINARLDAVSDV     558
human       TLRNLEILQNQTDMKTKGSLLWVLDHTKTSFGRRKLKKWVTQPLLKLREINARLDAVSEV     600
            ************************************:*********************:*

mouse       LHSESSVFEQIENLLRKLPDVERGLCSIYHKKCSTQEFFLIVKSLCQLKSELQALMPAVN     618
human       LHSESSVFGQIENHLRKLPDIERGLCSIYHKKCSTQEFFLIVKTLYHLKSEFQAIIPAVN     660
            ******** **** ******:**********************:* :****:**::****

mouse       SHVQSDLLRALIVEAPELLSPVEHYLKVLNGPAAKVGDKTELFKDLSDFPLIKKRKNEIQ     678
human       SHIQSDLLRTVILEIPELLSPVEHYLKILNEQAAKVGDKTELFKDLSDFPLIKKRKDEIQ     720
            **:******::*:* ************:**  ************************:***

mouse       EVIHSIQMRLQEFRKILKLPSLQYVTVSGQEFMIEIKNSAVSCIPADWVKVGSTKAVSRF     738
human       GVIDEIRMHLQEIRKILKNPSAQYVTVSGQEFMIEIKNSAVSCIPTDWVKVGSTKAVSRF     780
             **..*:*:***:***** ** ************************:**************

mouse       HPPFIVESYRRLNQLREQLVLDCNAEWLGFLENFGEHYHTLCKAVDHLATVDCIFSLAKV     798
human       HSPFIVENYRHLNQLREQLVLDCSAEWLDFLEKFSEHYHSLCKAVHHLATVDCIFSLAKV     840
            * *****.**:*************.****.***:*.****:*****.***************

mouse       AKQGNYCRPTLQEEKKIIIKNGRHPMIDVLLGEQDQFVPNSTSLSQDSERVMIITGPNMG     858
human       AKQGDYCRPTVQEERKIVIKNGRHPVIDVLLGEQDQYVPNNTDLSEDSERVMIITGPNMG     900
            ****:*****:***:**:********:**********:***.*.**:***************

mouse       GKSSYIKQVALVTIMAQIGSYVPAEEATIGIVDGIFTRMGAADNIYKGRSTFMEELTDTA     918
human       GKSSYIKQVALITIMAQIGSYVPAEEATIGIVDGIFTRMGAADNIYKGQSTFMEELTDTA     960
            ***********:****************************************:***********

mouse       EIIRRASPQSLVILDELGRGTSTHDGIAIAYATLEYFIRDVKSLTLFVTHYPPVCELEKC     978
human       EIIRKATSQSLVILDELGRGTSTHDGIAIAYATLEYFIRDVKSLTLFVTHYPPVCELEKN     1020
            ****:*: **************************************************

mouse       YPEQVGNYHMGFLVNEDESKQDSGDMEQMPDSVTFLYQITRGIAARSYGLNVAKLADVPR     1038
human       YSHQVGNYHMGFLVSEDESKLDPGAAEQVPDFVTFLYQITRGIAARSYGLNVAKLADVPG     1080
            * .***********.***** * *  **:** ***************************

mouse       EVLQKAAHKSKELEGLVSLRRKRLECFTDLWTTHSVKDLHTWADKLEMEEIQTSLPH        1095
human       EILKKAAHKSKELEGLINTKRKRLKYFAKLWTMHNAQDLQKWTEEFNMEETQTSLLH        1137
            *:*:*************:. :****: *:.*** *..:**:.*:::::*** **** *
```

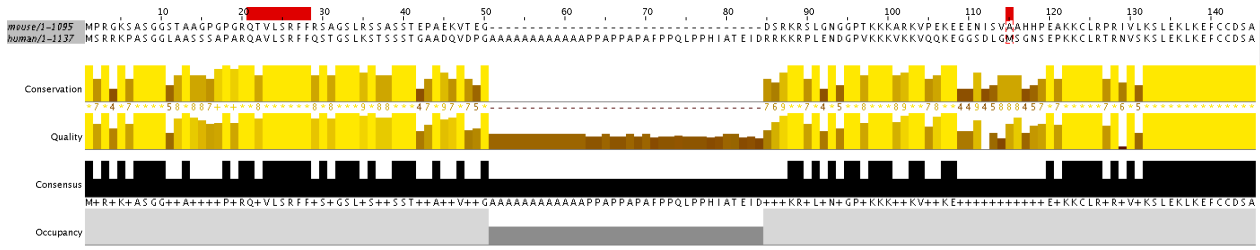


Figure 1.9. Conservation of the MSH3 N-terminal domain between mouse and human.
The PCNA interaction motif and Tome et al. (2013a) p.A82S variant are marked in red. Figure prepared in Jalview 2.10.4b1.

1.5.3 MSH3 depletion

Deficient MMR results in a mutator phenotype known as microsatellite instability (MSI), with length alterations at simple repeated sequences called microsatellites, which is the hallmark of Lynch syndrome, also known as hereditary non-polyposis colorectal cancer (HNPCC) (Yamamoto and Imai, 2015). It is usually caused by mutations in MLH1 and MSH2, and less frequently in MSH6 and PMS2. Mutation of *MSH3* has not been linked to cancer in humans, most likely because MutSα can also initiate repair at most replication errors (Edelmann et al., 2000, Jiricny, 2006), but loss of *MSH3* is associated with increased microsatellite instability (Haugen et al., 2008), and Msh3 knockout further increases cancer susceptibility in Msh6 knockout mice (de Wind et al., 1999, Edelmann et al., 2000).

1.6 The immune system in neurodegenerative disease

1.6.1 Neurodegenerative disease

Neurodegenerative diseases are characterised by synaptic loss and neuronal death resulting in cognitive decline and loss of motor function. These are attributed to aggregation of the pathogenic protein, which can occur spontaneously or due to inherited mutation. The diseases can be histologically classified by the pathological protein aggregate, which in amyloidoses such as CJD or Alzheimer's disease is the prion protein or plaques of Aβ, in the tauopathies there are neurofibrillary tangles of the hyperphosphorylated microtubule-binding protein tau, also present in AD, in the synucleinopathies such as Parkinson's disease aggregates of α-synuclein form as Lewy bodies, and aggregates of TDP-43 form in amyotrophic lateral sclerosis (Dugger and Dickson, 2017). In aging and neurodegeneration there is evidence for increased number and activation of microglia in the CNS (Srinivasan et al., 2016). High levels of proinflammatory cytokines, including TNF, IL-1β and IL-6 are seen in brain, CSF and serum of AD, PD and HD patients (Heneka et al., 2014), which are thought to derive from microglia, rather than infiltrating adaptive immune cells (Crotti and Glass, 2015).

1.6.2 Huntington's disease

The innate immune system shows prominent deficits in Huntington's diseases. Microglia are activated in the brains of HD patient and mouse models, even before symptom onset (Tai et al., 2007b, Simmons et al., 2007), and complement components C1q, C4 and C3 are upregulated in the striatum (Singhrao et al., 1999). Several proinflammatory cytokines are increased in peripheral blood plasma and in the brain of HD patients, including IL-6 and IL-8 (Bjorkqvist et al., 2008), potentially driven by the upregulation of NFκB (Khoshnan et al., 2004). Reduction of IL-6 levels by antibody neutralisation or CB2 agonism in HD mice extends life span and suppresses motor deficits, synapse loss, and CNS inflammation (Bouchard et al., 2012a). α2-macroglobulin (α2M), an acute phase protein, is also increased in HD patient brain, mainly in reactive astrocytes (Dalrymple et al., 2007, Du et al., 1998). HD patient and mouse innate immune cells, including monocytes, macrophages and microglia, show impaired migration to an inflammatory stimulus (Kwan et al., 2012b),

which may be related to the increased IL-1β seen in brain and serum of HD patients and mice, even before symptom onset; knockout of the chemokine receptor *Ccr2* in mice increases serum IL-1β and IL-6, leading to failure of macrophage and lymphocyte recruitment to inflammatory stimuli (Kurihara et al., 1997, Christensen et al., 2004, Rampersad et al., 2011).

Within the adaptive immune system, IL-4, an anti-inflammatory cytokine involved in tolerance and induction of regulatory T cells, is increased in HD patient and mouse brain and plasma (Bjorkqvist et al., 2008). Taken together, these results provide compelling evidence for immune upregulation, particularly of the innate immune system, in HD (Ellrichmann et al., 2013).

1.6.3 Summary

In summary, innate immune cells such as microglia can have both beneficial and damaging roles in the brain. We have currently only scratched the surface of the signalling pathways involved, but the study of neuroinflammation will be vital in understanding the pathogenesis of neurodegenerative disease. Risk genes have been identified relating to complement and microglial receptors, demonstrating a role for the immune system in AD pathogenesis. However, immune activity is complicated, with complement and peripheral immune cells having both beneficial and deleterious effects. While many immune components contribute to pathogenesis, it is unclear if they act cooperatively or independently. Innate immune pathways are amenable to pharmacological manipulation, so there is hope that a better understanding of the contribution of immune cells could lead to targeted therapies in the future.

Chapter 2 Materials and methods

2.1 Cell lines

2.1.1 HEK 293

Human embryonic kidney cells derived from a healthy foetus in 1973 and transformed by adenovirus.

2.1.2 SH-SY5Y

Derived from a bone marrow biopsy from a four-year-old girl with neuroblastoma.

2.1.3 HeLa

Derived from cervical cancer cells from a patient named Henrietta Lacks in 1951.

2.1.4 ReNcell neural stem cells

The human neural ReNcell VM and CX progenitor cell lines were derived from 10-week-old foetal ventral mesencephalon and cerebral cortex, and immortalised by retroviral transduction with the v-myc oncogene by ReNeuron Group PLC (Guildford, UK). The CX line is clonal. The cell lines are commercially available from Millipore (cat #SCC008 and SCC007). They have a stable karyotype and can differentiate into neuronal and glial cells (Millipore, 2016, Donato et al., 2007). They grow as a monolayer with a doubling time of 20-30 hours.

2.1.4.1 Lentiviral transduction with HTT exon 1

ReNcell neural stem cells were transduced with *HTT* exon 1 containing 29, 71 or 129 CAG repeats in the Tabrizi lab. The p'HRsincpptUCOE+htt exon 1 IRES eGFP vector was generated by modifying A2UCOE, from Zhang et al. (2007), as described in Trager et al. (2014). Exon 1 human HTT-IRES-eGFP was ligated into the SalI-NdeI sites. The internal ribosome entry site (IRES) permits expression of both HTT and GFP from a single vector. Vector sequences are given in the Appendix. Transduced cells were FACS sorted (fluorescence-activated cell sorting) by GFP expression, which is contained within the A2UCOE plasmid.

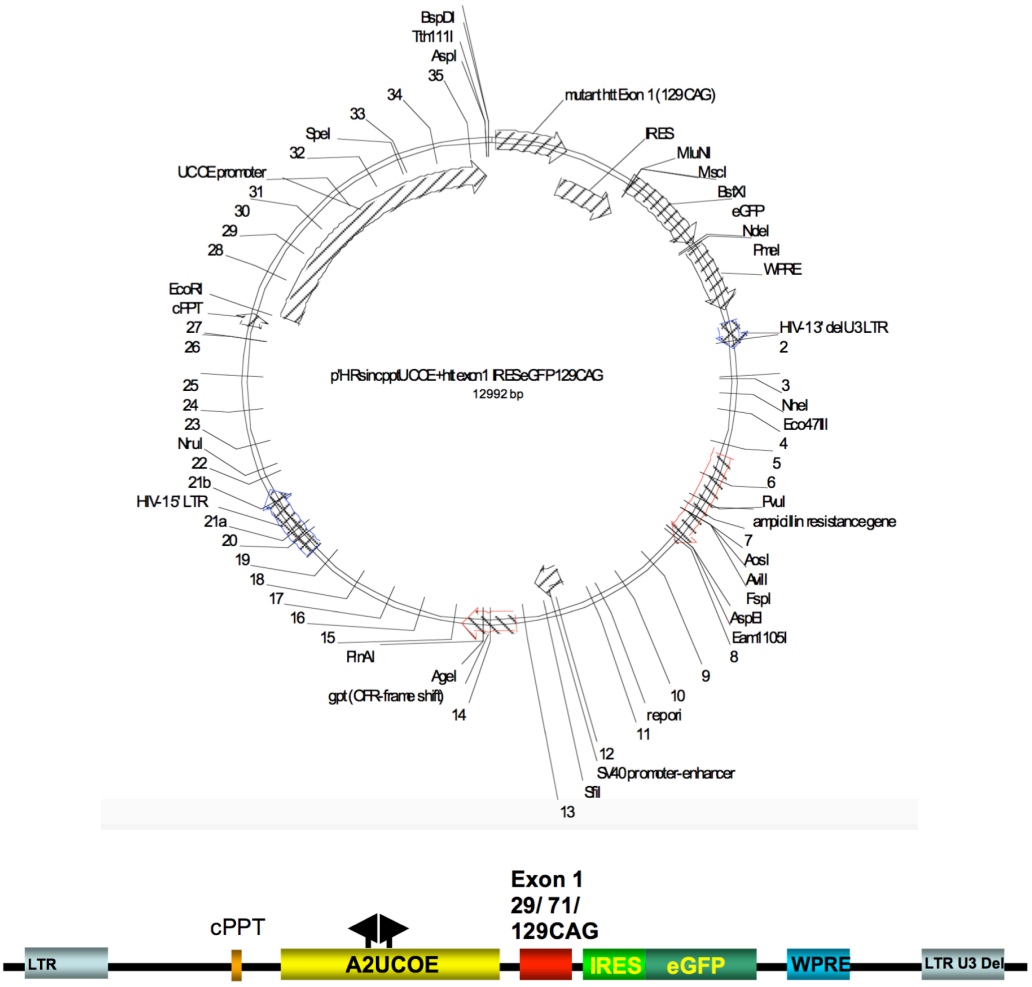


Figure 2.1. p'HRsincpptUCOE+htt exon1 IRES eGFP 129CAG vector.
Top – vector map showing sites of the exon 1 insertion, GFP and the ampicillin resistance gene. Below – linear representation.

Expression of the 29, 71 and 129 CAG *HTT* exon 1 construct was demonstrated by fluorescence of GFP, which is expressed from the cassette through an IRES, and by western blot using antibodies to HTT.

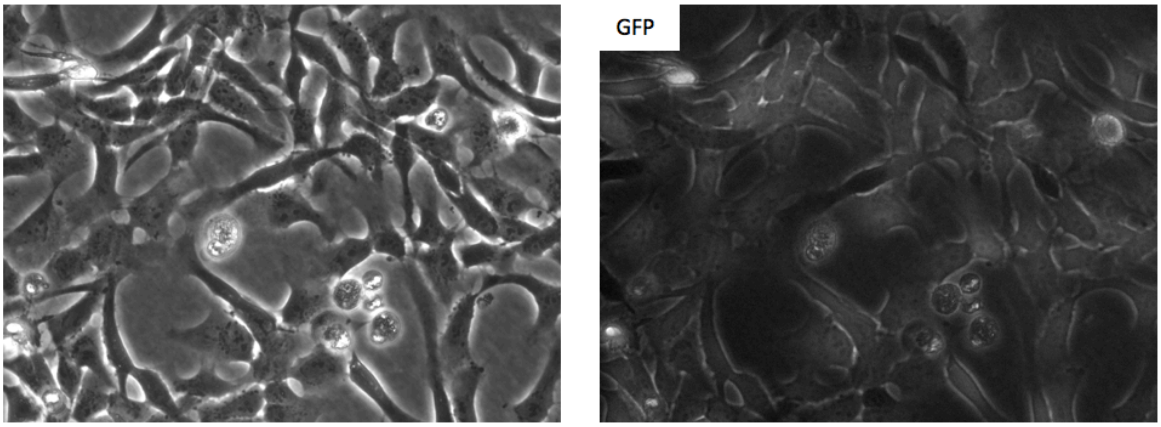


Figure 2.2. Micrographs of ReN VM cells transduced to express HTT exon 1 with 129 CAG repeats and GFP.
Left – 40x magnification. Right – GFP fluorescence.

Neuronal ReN VM cells transduced to express *HTT* exon 1 with 71 or 129 CAG repeats show HD-relevant phenotypes, including HTT aggregates and mitochondrial respiratory chain deficits. Our group has conducted a baseline analysis of aggregate formation, size and location using a panel of validated antibodies to HTT, including S830 and MAB5492 (Millipore), on the UCL Perkin Elmer Opera automated microscopy system (unpublished).

2.1.4.2 Single cell cloning

Cells were serially diluted and cultured in 20 µg/ml laminin-coated 96 well plates at either 1000, 100 or 10 cells/well in 200 µL of NSC media (see below). Media was changed twice weekly.

2.1.5 Track-HD patient-derived cell lines

In the Track-HD cohort (Tabrizi et al., 2013, Tabrizi et al., 2012, Tabrizi et al., 2011a, Tabrizi et al., 2009a) our group developed a progression score that incorporated longitudinal imaging, cognitive and quantitative motor measures, and controlled for CAG repeat length (Hensman Moss et al., 2017b). Whole exome sequencing in the 25 fastest and 23 slowest progressing subjects identified several coding variants in *FAN1* which were candidates for functional analysis. p.R507H (rs150393409), is a relatively rare SNP (MAF = 0.0028) in the DNA binding domain which was the third most significant variant in the GeM GWAS of HD, is associated with a 5.5 year early onset (p = 9.34E-18) (GeM-HD, 2015) and is predicted damaging *in silico* (SIFT and polyphen). It was observed in two fast progressing Track-HD subjects with onset 7.22 and 10.31 years earlier than expected (mean 8.76 $\pm$ 1.54 years, progression scores 1.83 and 0.42 respectively). Their mean progression score was 1.12 ($\pm$ 0.70), which is equivalent to an acceleration of up to 1.30 units on the UHDRS total motor score (TMS) and 0.37 units on the total functional capacity (TFC) per year (Hensman Moss et al., 2017b). Other variants include p.R145H (rs146408181, 10.1y early onset, progression score 2.11), p.E240K (rs150748572, 10.2y early onset, progression score 2.36) and p.829H (rs781332818, 9.3y early onset, progression score 2.27). In addition, two rare coding variants associated with slow progression were found; p.P894S (rs80120912), in the nuclease domain, was found in two subjects (progression scores -0.76 and -1.59), and p.R377W in one (rs151322829, progression score -1.25).

SNP id	Ch:location (GRCh38.p7)	Minor allele frequency (1000G)	Consequence	SIFT	Polyphen	n of het subjects in TrackHD	Protein location	Reference amino acid	Alternative amino acid	Domain	Slope in GeM GWAS (years/minor allele)	P-value in GeM GWAS	Mean progression score (+/- sd)	Mean residual AAO (years +/- sd)
rs146408181	15:30905097	0.0002	missense variant	tolerated	benign	1	145	P	H	-	-	-	2.11	-10.1
rs150748572	15:30905381	0.0012	missense variant	tolerated	benign	1	240	E	K	-	-	-	2.36	-10.2
rs151322829	15:30905792	0.0014	missense variant	deleterious	benign	1	377	R	W	-	-	-	-1.25	-
rs150393409	15:30910758	0.0028	missense variant	deleterious	possibly damaging	2	507	R	H	SAP	-5.55	9.34E-18	1.12 (+/- 0.70)	-8.76 (+/- 1.54)
rs781332818	15:30925938	0.0000	missense variant, splice region	deleterious	possibly damaging	1	829	Q	H	-	-	-	2.27	-9.3
rs80120912	15:30929290	0.0080	missense variant	tolerated	benign	2	894	P	S	VRR NUC	-	-	-1.17 (+/- 0.42)	1.5

Table 2.1. FAN1 variants identified by whole exome sequencing (WES) in fast and slow progressing subjects from TRACK-HD.
SIFT and polyphen represent in silico functional prediction.

Group	Subject iD	FAN1 variant	Stage	Predicted AAO	Age at baseline	AAO	CAG	Progression score	Gender	Control for
Cases	326-549-639	p.R145H	Manifest HD	48.05	37.27	38	43	2.11	Male	-
	850-476-675	p.E240K	Manifest HD	52.22	44.16	42	42	2.36	Male	-
	598-074-99X	p.R377W	Manifest HD	62.62	58.45	-	40	-1.25	Male	-
	432-196-753	p.R507H	Manifest HD	52.22	44.73	45	42	1.83	Female	-
	135-074-011	p.Q829H	Manifest HD	41.34	35.75	32	45	2.27	Female	-
	826-350-503	p.P894S	Premanifest	57.04	49.40	-	41	-1.59	Male	-
	535-845-735	p.P894S	Manifest HD	62.62	64.13	-	40	-0.76	Male	-
Controls	841-795-617	-	Premanifest	48.05	37.34	-	43	-1.71	Male	326-549-639 (p.R145H)
	147-577-839	-	Premanifest	52.22	43.93	-	42	-1.52	Male	850-476-675 (p.E240K) and 432-196-753 (p.R507H)
	932-487-54X	-	Premanifest	52.22	53.17	-	42	-2.44	Female	432-196-753 (p.R507H)
	768-309-063	-	Premanifest	41.34	34.96	-	45	-1.03	Female	135-074-011 (p.Q829H)

Table 2.2. Track-HD patient-derived lymphoblastoid (LB) cell lines used in this study.
AAO – age at motor onset, CAG – pathogenic HTT CAG repeat length.

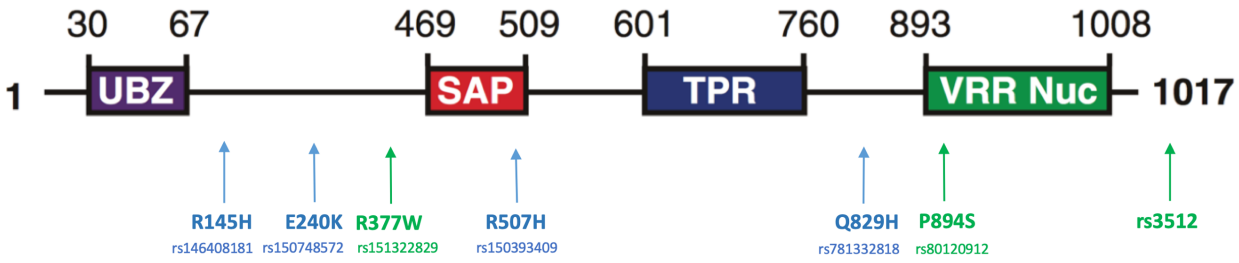


Figure 2.3. Schematic representation of FAN1.
Blue text – variants associated with fast progression, green text – variants associated with slow progression. The UBZ (ubiquitin-binding zinc finger 4), SAP (SAF-A/B, Acinus and PIAS), TRP (tetratricopeptide repeat) and VRR Nuc (viral replication and repair nuclease) domains of FAN1 are indicated.

Chapter 3 identifies *FAN1* variants that also modify onset in other polyglutamine diseases (Bettencourt et al., 2016). rs3512 (g.30942802G>C), a common variant (MAF 0.19) in the 3' untranslated region (UTR) of *FAN1*, which was associated with 1.3-year delayed onset in the GeM GWAS (p=5.28E-13) (GeM-HD, 2015), is associated with a 1.7-year delayed onset in the polyglutamine diseases (p = 1.52E10-05). rs3512 is in high LD (r^2 = 0.85) with, rs2140734, the HD GWAS secondary peak at the *FAN1* locus which was associated with 1.4-year late onset (p = 7.1E-14). Within the TRACK-HD cohort there are 18 homozygotes for the minor allele at rs3512, five of whom are amongst the slowest progressors. Skin biopsies have been collected from two of these.

2.1.6 250Q lymphoblasts

Dr Lara Cravo (Oxford University) shared lymphoblasts generated from a juvenile-onset HD subject with a 250 CAG repeat expansion (Nance et al., 1999). The subject had clinical onset aged 2.5 years with rigidity and cognitive decline, followed at 5 with seizures. He lived to 16, at which point he was mute with joint contractures and little spontaneous movement. His mother, from whom he likely inherited the mutation, died at 31 but had not been diagnosed with HD. Original sizing was based on a long smear on the PCR gel that was centred around 250 CAG.

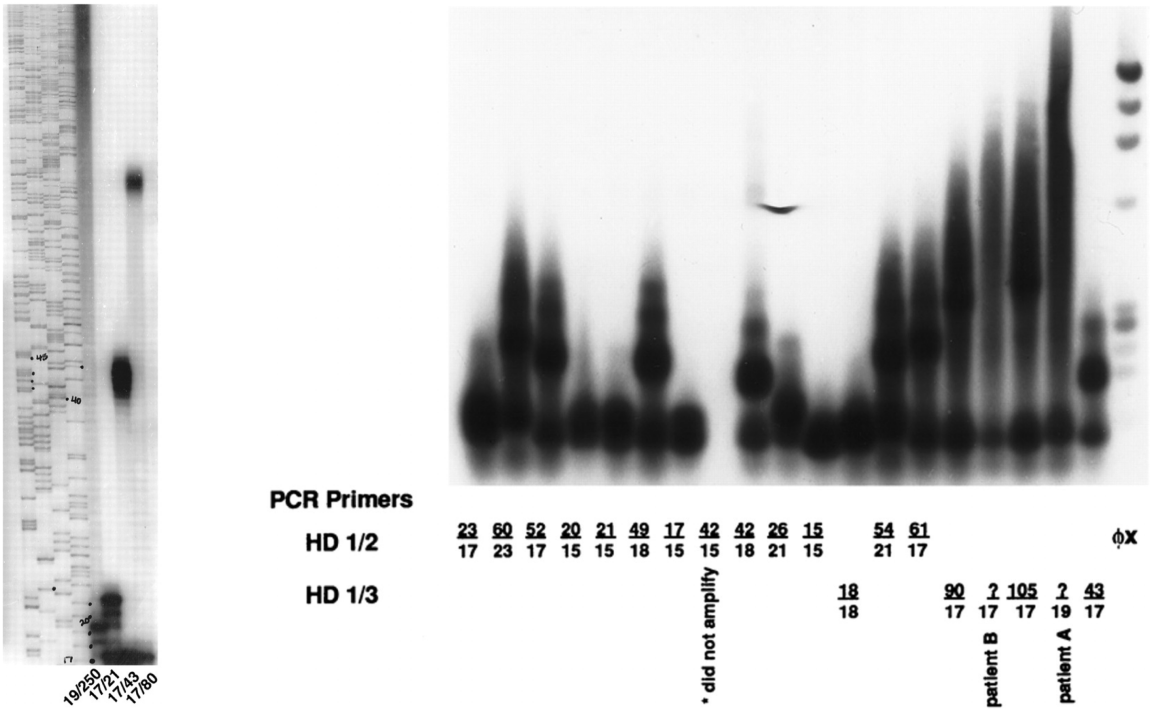


Figure 2.4. CAG repeat sizing in 250Q lymphoblasts from Nance et al. (1999).
Left – PCR amplification of genomic DNA. The 250Q subject is indicated by 19/250. The expanded allele appears as a broad smear on denaturing polyacrylamide (5%) gel. Right – Southern transfer, the 250Q subject is indicated by 'patient A'. The densest region of the smear corresponds to approximately 250 CAG.

2.1.7 73Q induced pluripotent stem cells

Skin biopsies were collected at UCLH from three juvenile-onset HD siblings with 56, 67 and 73 CAG repeats, as well as their mother, an unaffected control, and used to generate fibroblasts. induced pluripotent stem cells (iPSC) were generated by Sendai virus reprogramming with the CytoTune-iPS 2.0 Sendai reprogramming kit by Prof Ali Brivanlou (Rockefeller University). iPSCs were verified by the expression of pluripotency markers, differentiation into all germ layers using a self-organisation assay, karyotyping, Sanger sequencing to confirm the CAG repeat length, and confirmation of the absence of exogenous Sendai virus. Three iPSC clones were generated from each subject.

2.1.8 125Q juvenile-onset HD patient-derived cells

2.1.8.1 Lymphoblastoid cells

Blood was taken separately for the generation of lymphoblastoid cells and induced pluripotent stem cells. For the former, three CPDA yellow top blood tubes were drawn on 2/12/16 and sent at room temperature the same day to the European Collection of Authenticated Cell Cultures (ECACC) lab at Public Health England (PHE), Salisbury, where peripheral blood lymphocytes (PBL) were isolated and frozen ahead of Epstein–Barr virus (EBV) transformation. Once the cell line was generated, four ampoules were prepared and frozen in liquid nitrogen. Samples were quality assured by mycoplasma screening, sterility, cell count and viability against the source material (a blood spot card prepared at receipt) using STR-PCR (short tandem repeat) profiling.

2.1.8.2 Induced pluripotent stem cells (iPSC)

4 vials of 6 ml whole blood were drawn on 19/9/17 in a BD Vacutainer cell preparation tube with sodium citrate (cat #362782). Two were shipped to Censo Biotechnologies, Midlothian, at room temperature the same day. The other two

vials were used to isolate peripheral blood mononuclear cells (PBMC) at UCL, which were stored in our lab. For the latter, the tube was mixed by gently inverting 8-10 times, then centrifuged for 30 min at 1700 xg at room temperature. After centrifugation, mononuclear cells and platelets were in a whitish layer just under the plasma layer. The entire contents of each tube above the gel were combined by pipetting into a 50 ml tube. PBS was added to bring the volume to 15 ml per vacutainer tube added i.e. 30 ml. Cells were mixed by inverting 5 times, then centrifuged at 300 xg for 15 min at room temperature. As much supernatant as possible was aspirated without disturbing the cell pellet, then 10 ml PBS was added, and the cells were resuspended by gently inverting the tube 5 times. Cells were centrifuged at 300 xg for 10 min at room temperature, then the supernatant was again aspirated without disturbing the cell pellet.

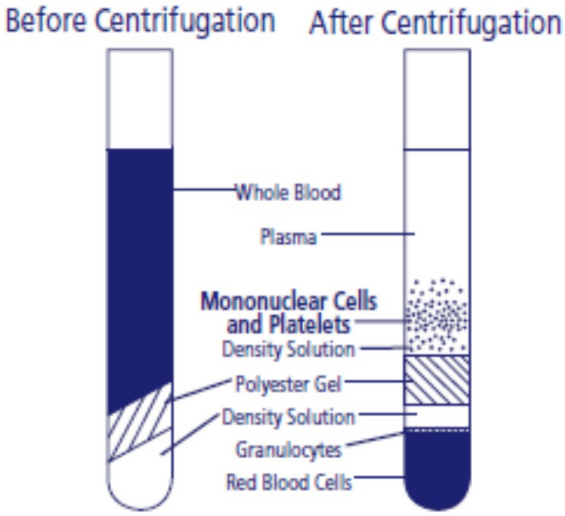


Figure 2.5. BD vacutainer centrifugation.

Freeze mix was prepared by adding 500 µL DMSO to 4.5 ml FBS and prechilled to 4°C. Cells were resuspended in 3 ml of freeze mix at 5.6E06 cells/ml, and 1 ml of cell suspension was added to each of 3 cryovials. These were placed in a Mr. Frosty freezing tub overnight at -80°C, then transferred to liquid nitrogen within 24 hours (PDG: 57399). One vial was subsequently shipped to Censo on 11/10/17, as cells from the fresh blood samples had been slow to establish.

Censo reprogrammed PBMCs to pluripotent erythroid progenitor cells (EPCs) using the Sendai cytotune 2.0 method. Cells were cultured on Matrigel in mTESR^{TM}1 medium, passaged using EDTA, and cryopreserved in Cryostor CS10 medium. They were quality assured for sterility, including mycoplasma, HIV-1, HIV-2, HBV, HCV and microbiological growth, were viable as evidenced by growth to confluency in <5 days, were karyotypically stable with no major abnormalities over 100 kb detected on array-based Comparative Genomic Hybridization (aCGH), and qPCR for the Sendai backbone showed clearance of the viral vector. The cells had a typical morphology and consistent 4-5 day growth cycle. The line is capable of spontaneous differentiation with reduction of self-renewal markers (differentiation index -2.48) and a clear increase in germ layer gene expression for mesoderm (differentiation index 6.12) and endoderm (differentiation index 2.44), with ectoderm showing a small but positive shift in gene upregulation (differentiation index 0.33), likely indicating potential for ectodermal differentiation in a longer spontaneous or directed differentiation assay. Marker expression was consistent with morphology, with the majority of self-renewal markers showing expression levels typical of human iPSCs (TRA-1-60 94.58% positive, POU5F1 99.44% positive, SSEA-4 99.92% positive, SSEA-3 48.46% positive, NANOG 84.18%

positive, SSEA-1 3.48% positive). Though NANOG and SSEA-3 were low, the normal expression of other markers in conjunction with typical morphology and the line's ability to differentiate, indicates a phenotype consistent with a pluripotent stem cell line. In summary, the expression profile and differentiation potential were typical of a human pluripotent stem cell line. 6 vials, each containing 1.0E06 cells at passage 9, were returned.

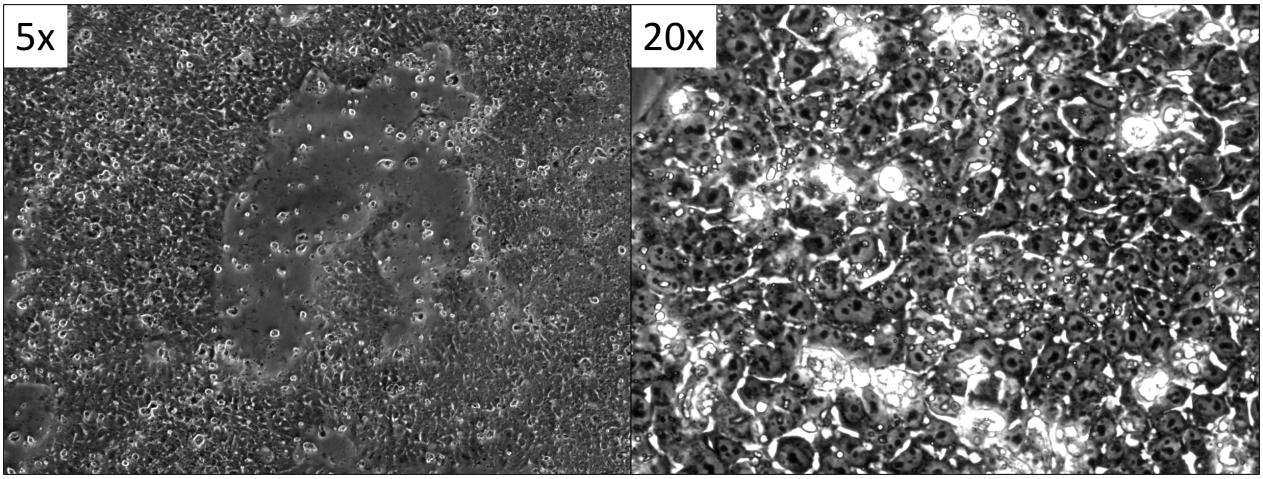


Figure 2.6. Micrographs of 125Q pluripotent ESCs.

2.1.9 109Q induced pluripotent stem cells (iPSC)

Fibroblasts from a female juvenile-onset HD subject with 19/109 CAG repeats (Coriell, ND39258) originated from Johns Hopkins University under Russell Margolis's IRB protocol #NA00018358. The iPSC line, gifted from Prof Nick Allen, Cardiff University, was generated at the Cedars-Sinai Medical Center and named CS09iHD109. Fibroblasts were reprogrammed into iPSCs by a non-integrating and virus-free method using the Amaxa Human Dermal Fibroblast Nucleofector Kit to express episomal plasmids with six factors: OCT4, SOX2, KLF4, L-MYC, LIN28 and p53 shRNA. Cedars-Sinai iPSC Core characterised the iPSCs, showing a normal karyotype, performed PCR to confirm the absence of episomal plasmids, and demonstrated pluripotency by immunostaining, RT-qPCR for endogenous pluripotency genes, gene-chip and bioinformatic PluriTest assays, and spontaneous embryoid body differentiation confirming the capacity to form all germ layers (Mattis et al., 2015, Consortium, 2017, Grima et al., 2017, Wiatr et al., 2018).

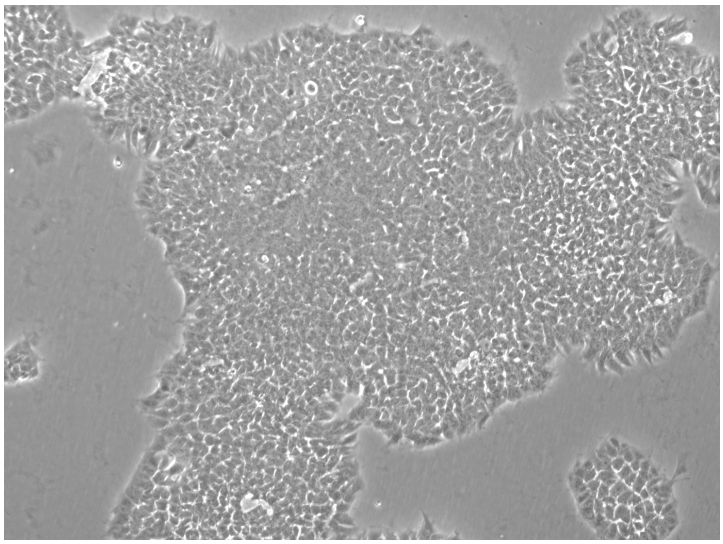


Figure 2.7. Representative light micrograph of 109Q iPSCs (5x).

2.1.10 U20S Flp-In

Munoz et al. (2014) shared the U20S Flp-In T-Rex osteosarcoma cell line in which endogenous *FAN1* has been knocked out by a novel Cas9/CRISPR technique. Briefly, cells were transfected with a vector containing an FRT Flp recombinase target site, an ATG start codon and a zeocin resistance gene. They were then transfected with the pcDNA5/FRT/TO vector that contains tetracycline-inducible Cas9 nuclease and a hygromycin resistance gene lacking an ATG initiation site and embedded in an FRT site.

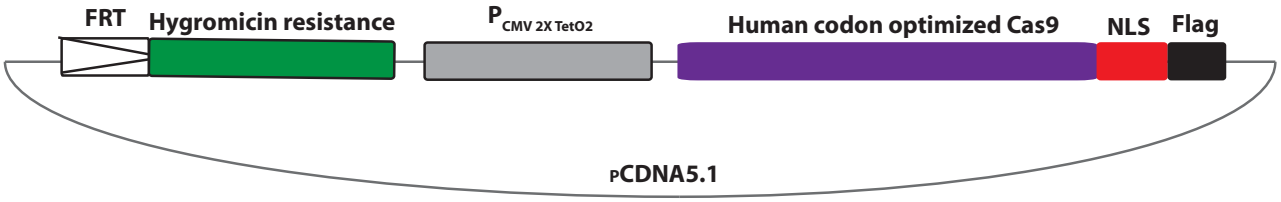


Figure 2.8. pcDNA5/FRT/TO vector containing Cas9-Flag under a tetracycline-inducible promoter and a hygromycin resistance gene.
Reproduced from Munoz et al. (2014)

Transfection with pOG44, a Flp recombinase, leads to recombination between FRT sites which is conservative, preserving both sites. This inactivates the zeocin resistance gene, generating stable Cas9 expression and hygromycin resistance.

Cells were transfected with sgRNA (single-guide RNA) which binds to its complementary sequence in *FAN1*. Cas9 is recruited to the sgRNA scaffold domain inducing a double strand break and non-homologous end joining leads to deletions and insertions, thereby disrupting *FAN1*. The Cas9 was then removed (flipped out) by reintroducing pOG44, which leads to recombination between the FRT sites. This disrupts hygromycin resistance and restores zeocin resistance.

Cas9 was then removed from the pcDNA5/FRT/TO vector and replaced with GFP-labelled full length *FAN1*, into which variants can be inserted by mutagenesis. This allows the *FAN1* knockout cells to be complemented with tetracycline-inducible GFP-FAN1 containing our variants of interest. These cells, which have the advantage of an isogenic background lacking the endogenous wild type *FAN1* allele, permit the study of FAN1 function. A line expressing endogenous FAN1 was provided as a control ($FAN1^{+/+}$).

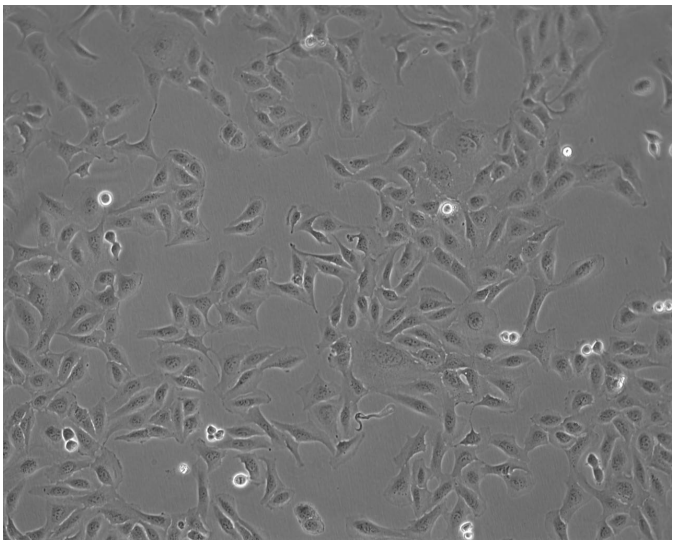


Figure 2.9. Light micrograph of U20S FAN$^{-/-}$ cells in culture.

2.2 Cell culture

2.2.1 Lymphoblastoid cells

LB cells were cultured in RPMI medium (Thermo, cat #21870-076) supplemented with 15% non-heat inactivated fetal bovine serum (FBS), 100 U/ml penicillin and 100 µg/ml streptomycin.

2.2.2 ReNcell

2.2.2.1 Routine culture

Cell proliferation is maintained in the presence of growth factors bFGF and EGF and their withdrawal results in differentiation into oligodendrocytes, astrocytes, and neurons within few days (Hoffrogge et al., 2006, Donato et al., 2007). Undifferentiated cells were maintained on Thermo Nunc plasicware coated with 10 µg/ml laminin in NSC media; DMEM:F-12 (Gibco, #21331046) with the following additives (Wood-Kaczmar et al., 2008). Media was changed every 48-72h.

- 0.03% human albumin solution (Sigma # A9080)
- 5 µg/ml transferrin (Sigma #T0665)
- 16.2 µg/ml putrescine (Sigma #P5780)
- 5 µg/ml insulin (Sigma #I9278)
- 400 ng/ml L-thyroxine (T4, Sigma #T2376)
- 337 ng/ml tri-iodo-thyronine (T3, Sigma #T2877)
- 60 ng/ml progesterone (Sigma #P6149)
- 40 ng/ml sodium selenite (Sigma #S9133)
- 10 U/ml heparin (Sigma #H3149)
- 40 ng/ml corticosterone (Sigma #46148)
- 1x penicillin and streptomycin (Sigma #P4333)
- 2 mM glutamine (Life technologies #25030-024)
- 10 ng/ml bFGF (Peprotech, #100-18B)

- 20 ng/ml EGF (Peprotech, #AF-100-15).

2.2.2.2 Differentiation

Cells were seeded at 30,000 cells/cm^2 and expanded to 80% confluence on Thermo Nunc plasicware coated with 20 µg/ml laminin. Differentiation was initiated (day 0) by changing to *Initial differentiation medium* that lacked the growth factors FGF and EGF, but was supplemented with 0.5 mM dibutyryl-cAMP (Calbiochem, #28745) and 2 ng/ml GDNF (Peprotech, #450-10). On day 6, media was changed to *Mid-differentiation medium* that lacked cAMP and GDNF. On day 15 they were changed into *Differentiated medium*, with 0.5 mM glutamine. This protocol, well established in the Tabrizi lab, generates a pan-neuronal culture (Wood-Kaczmar et al., 2008).

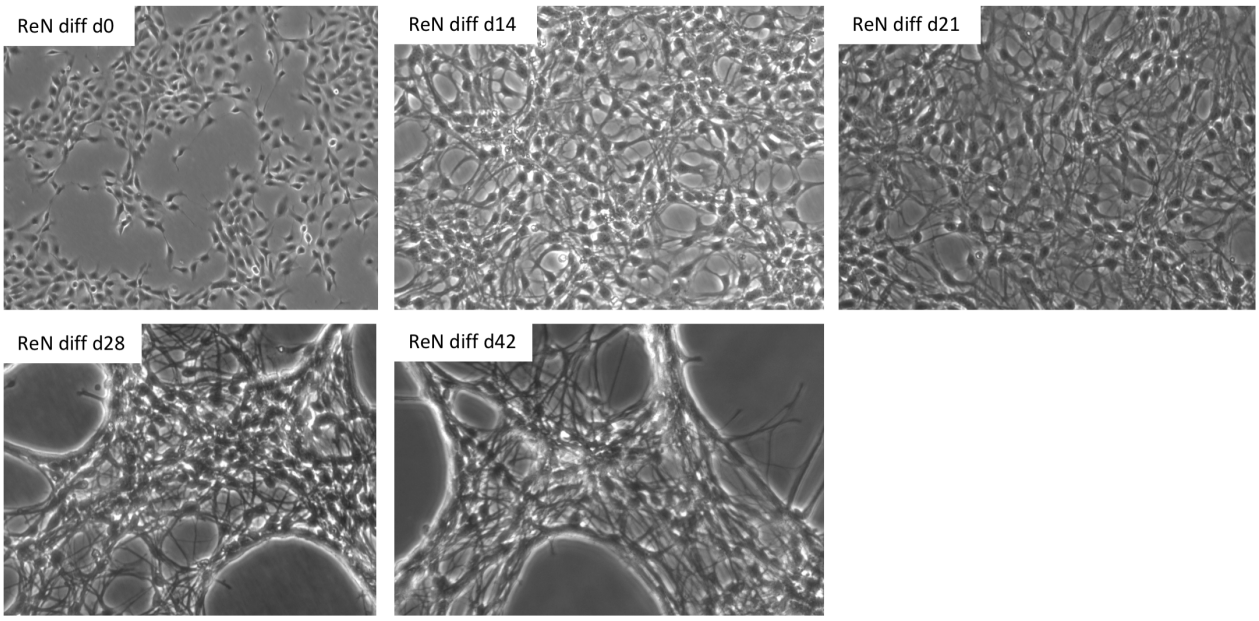


Figure 2.10. Neuronal differentiation of ReN VM cells expressing HTT exon 1 with 129 CAG repeats.
Light micrographs with days from initiation of differentiation shown.

ReN CX cells expressing 129 CAG repeats rarely successfully differentiated without losing expression of *HTT* exon 1 (1/10 differentiations). Representative micrographs from a sample differentiation are shown below.

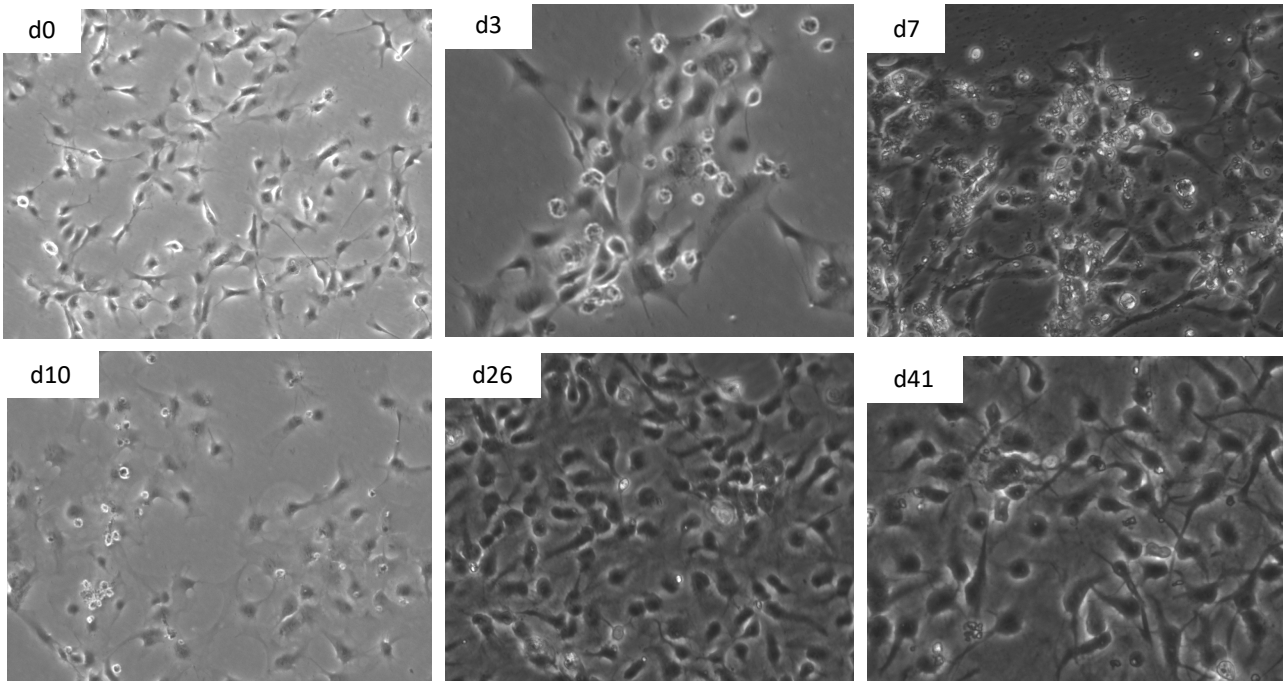


Figure 2.11. Neuronal differentiation of ReN CX cells expressing HTT exon 1 with 129 CAG repeats.
Light micrographs with days from initiation of differentiation shown.

2.2.3 Stem cells

2.2.3.1 Reagents

- DMEM/F-12 (w/o glutamine) — Gibco cat no. 21331-046
- Neurobasal — Gibco cat no. 21103-049
- D-PBS — Gibco cat no. 14190-094
- N2 — Gibco cat no. 17502-048
- B27 (w/o Vitamin A) — Gibco cat no. 12587-010
- B27 (with Vitamin A) — Gibco cat no. 17504-044
- L-Glutamine — Gibco cat no. 25030-024
- β-mercaptoethanol — Gibco cat no. 21985-023
- LDN193189 — Sigma cat no. SML0559
- SB431542 — Cambridge Biosciences SM33-10
- Dorsomorphin — Cambridge Biosciences SM03-10
- Activin A — PeproTech cat no. AF-120-14E
- Y27632, ROCK inhibitor — Sigma cat no. Y0503
- BDNF — PeproTech cat no. AF450-02
- GDNF — PeproTech cat no. AF-450-10
- Fibronectin — Millipore cat no. FC010
- Poly-D-lysine — Sigma cat no. P7280
- Laminin — Trevigen cat no. 3400-010-02
- Geltrex — Gibco cat no. A1413302

- Essential 8 medium — Gibco cat no. A1517001
- EDTA (0.5 mM) — Gibco cat no. 15575-020
- EGF — Peprotech cat no. AF-100-15
- bFGF — Peprotech cat no. 100-18B

2.2.3.2 Routine culture

This culture protocol was adapted from Shi et al. (2012). iPSCs were cultured on Geltrex-coated 6-well plates in E8+ media (10ml of 50X supplement added to 500 ml Essential 8 media, Thermo #A1517001). To coat the plates, 1 ml 1:100 Geltrex:cold DMEM:F12 was used per well of a 6-well plate, and plates were incubated at 37°C for a minimum of 1 hour prior to use. Cells were passaged at 80-85% confluency at a split ratio of 1:6. Cells were washed in D-PBS, 1 ml of 0.5 mM EDTA in D-PBS added and then incubated at 37°C for 3 min. EDTA was aspirated, 2 ml of E8+ added and cells were gently triturated. Cells were diluted and added to a new plate. Generally, cells needed passage every 4-5 days.

2.2.3.3 Thawing

One vial of frozen iPSCs was thawed into one well of a 6-well plate. The vial was removed from liquid nitrogen and immersed in a water bath at 37°C. Once thawed, cells were transferred into a 15 ml tube and 10 ml of room temperature E8+ added. Cells were centrifuged at 200 xg for 5 mins, the supernatant discarded, and the pellet gently resuspended in 2 ml of E8+. Geltrex solution was removed from the wells and the 2 ml iPSC suspension added dropwise around the edges. Media was changed daily until the first passage.

2.2.3.4 Freezing

To freeze iPSCs, the passage was conducted as above, but following aspiration of EDTA, 900 µL ice cold E8+ was added per well to triturate cells, the suspension was transferred to a cryovial, 100 µL DMSO added and the vial was frozen at -80°C in a Mr Frosty™ Freezing Container (Nalgene). 24 h later it was transferred to liquid nitrogen.

2.2.3.5 Medium spiny neuron (MSN) differentiation

This differentiation protocol was adapted from Arber et al. (2015). N2B27 differentiation medium was prepared using 2/3 DMEM:F12 and 1/3 Neurobasal, supplemented with 1:100 L-Glutamine, 1:150 N2, 1:150 B27 (without vitamin A up until day 26, and with vitamin A thereafter), and 0.1 mM β-mercaptoethanol. Cells were washed with D-PBS, then 2 ml of room temperature N2B27 differentiation medium, supplemented with SMAD inhibitors 100 nM LDN (1:5000), 10 µM SB431542 (1:1000) and 200 nM dorsomorphin (1:5000), was added. Half media was changed on alternate days.

Between day 9 and 12, cells were passaged at a ratio of 2:3 onto fibronectin-coated 12 well plate. Plates were coated with 0.5 ml per well of 25 µg/ml fibronectin solution in D-PBS (1:40), incubated at 37°C for at least 1 h and washed with D-PBS before use. An hour before passage, media was removed from cells and replaced with 1.5 ml N2B27 differentiation medium supplemented with 25 ng/ml (1:4000) activin A and 10 µM ROCK inhibitor (1:1000), and incubated at 37°C. After 1 h the conditioned media was reserved in a falcon tube and the cells washed in D-PBS. 0.5 ml 0.02% EDTA was added and cells incubated for 1 min at 37°C. EDTA was aspirated, 1 ml conditioned media added, and the surface scratched in pattern 1 with the tip of a 10 ml serological pipette to generate large clusters of cells. Cells were collected into a falcon tube and diluted with fresh N2B27 media with activin A and ROCK inhibitor, to an appropriate volume for a 2:3 split ratio. After removing PBS from the fibronectin-coated plates, 1.5 ml of the cell suspension was distributed dropwise around

the edge of each well. The following day 1 ml media was replaced with 1.5 ml N2B27 containing activin A but no ROCK inhibitor. Half media was changed on alternate days, thereby diluting out the ROCK inhibitor.

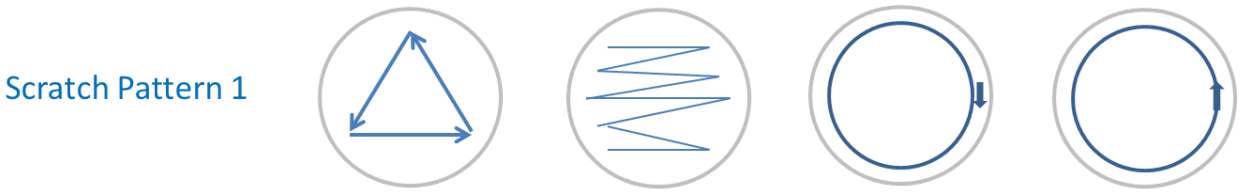


Figure 2.12. Scratch pattern 1 for MSN differentiation passage 1.

A second passage was conducted between days 19 and 22, at the neural progenitor stage when cultures were multi-layered with occasional rosette formation. Cells were passaged at a ratio between 1:1 and 1:4 onto poly-D-lysine/laminin coated 6 or 12 well plates, or coverslips. To coat, 1 ml of 0.01% poly-D-lysine in dH_2O was added per well of a 6-well plate (100 µL per coverslip) and the plate incubated at 37°C for at least 2 h. Plates were washed with sterile dH2O and 1 ml of 20 µg/ml laminin diluted in cold DMEM-F12 was added (100 µL per coverslip) and incubated at 37°C for at least 1h. Cells were washed in D-PBS, then 0.5 ml 0.02% EDTA was added and incubated for 1-2 min at 37°C. EDTA was aspirated and 1 ml of N2B27 medium with activin A was added to each well. The surface was scratched with the tip of a P1000 pipette tip, to generate small clusters of cells, as shown in scratch pattern 2. Cell clusters were collected into a falcon tube and resuspended by trituration with a 5 ml stripette. The cell suspension was diluted with an appropriate volume of N2B27 medium with activin A, and 2 ml added around the edge of each well of a 6-well plate. Half volume media changes were carried out with N2B27 medium with activin A on alternate days.

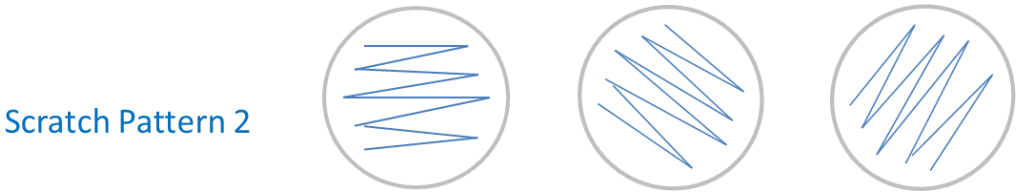


Figure 2.13. Scratch pattern 2 for MSN differentiation passage 2.

At day 26, half media was changed to N2B27 supplemented with activin A (25 ng/ml), 20 ng/ml BDNF and 20 ng/ml GDNF. Thereafter half media was changed on alternate days. Arber et al. (2015) reported that by day 35, 80% of cells are neuronal (NeuN+), of which 50% display CTIP2 staining and 20-50% are DARPP32 positive.

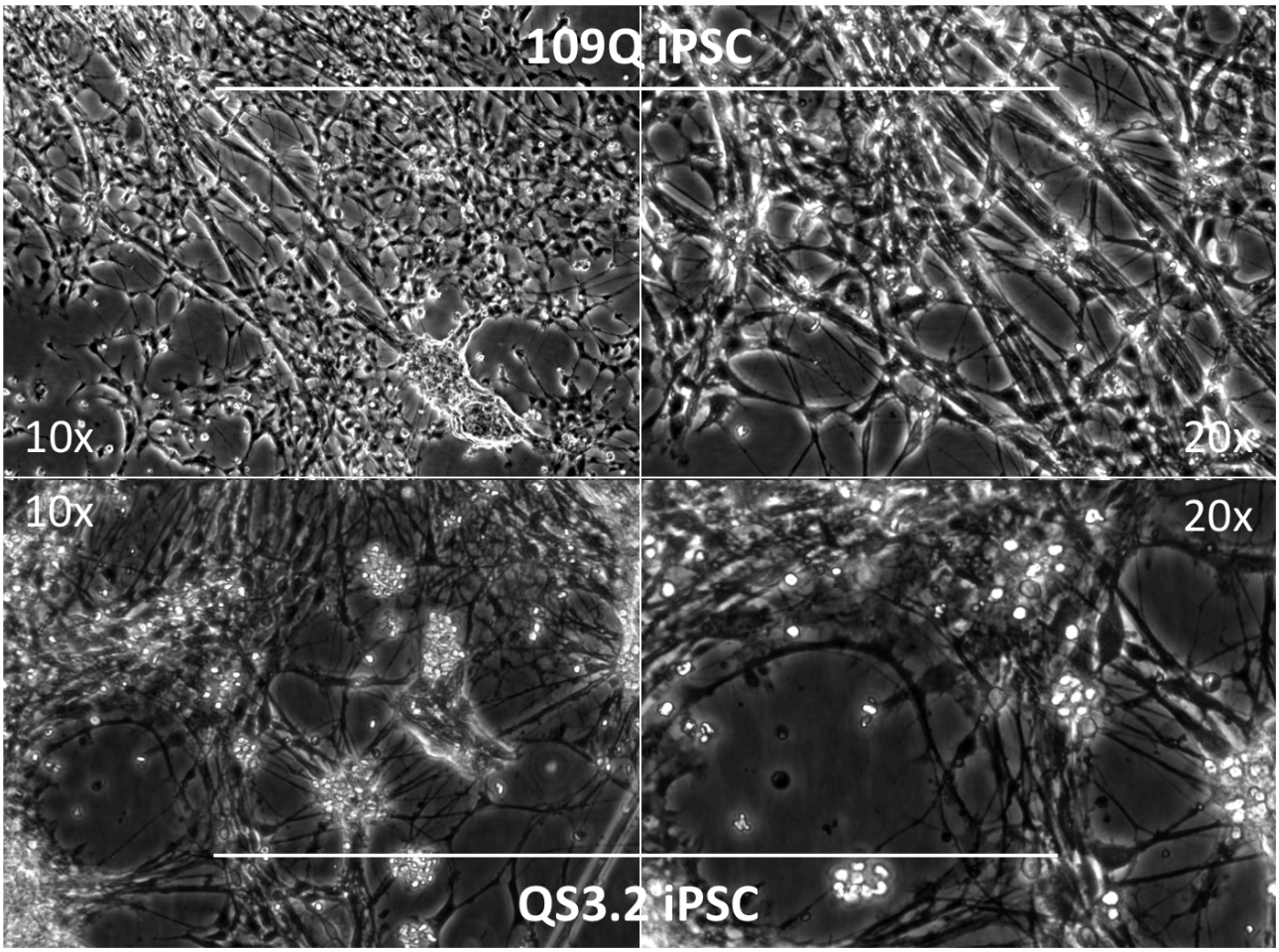


Figure 2.14. Light micrographs of differentiated medium spiny neurons (MSN).
***Top row** – 109Q MSNs, **bottom row** – 73Q clone 2 (QS3.2) MSNs, **left column** – 10x magnification, **right column** – 20x magnification.*

2.2.3.6 Maintenance of neural stem cells

Cultures can be maintained in the neural stem cell (NSC) stage at MSN differentiation passage 2 on day 19-20 (Shi et al., 2012). 6 well plates were coated with 20 µg/ml laminin in cold DMEM-F12 and incubated at 37°C for at least 1h. Cells were washed with D-PBS, then 0.5 ml Accutase was added and incubated at 37°C for 3-5 min. Cells were triturated gently with a P1000 pipette, transferred to a 15 ml falcon containing 10 ml D-PBS, then centrifuged at 300 xg for 3 min. The supernatant was removed, and the pellet resuspended in MSN NSC medium, prepared from N2B27 (without vitamin A) supplemented with 25 ng/ml activin A, 20 ng/ml bFGF and 20 ng/ml EGF. Having removed the laminin solution, 2 ml of cell solution was distributed into each well. Media was changed on alternate days and the cells passaged as required with either Trypzean or accutase. NSCs can re-enter MSN differentiation at the passage 2 stage by plating at 80% density on poly-D-lysine/laminin coated plates or coverslips, as above.

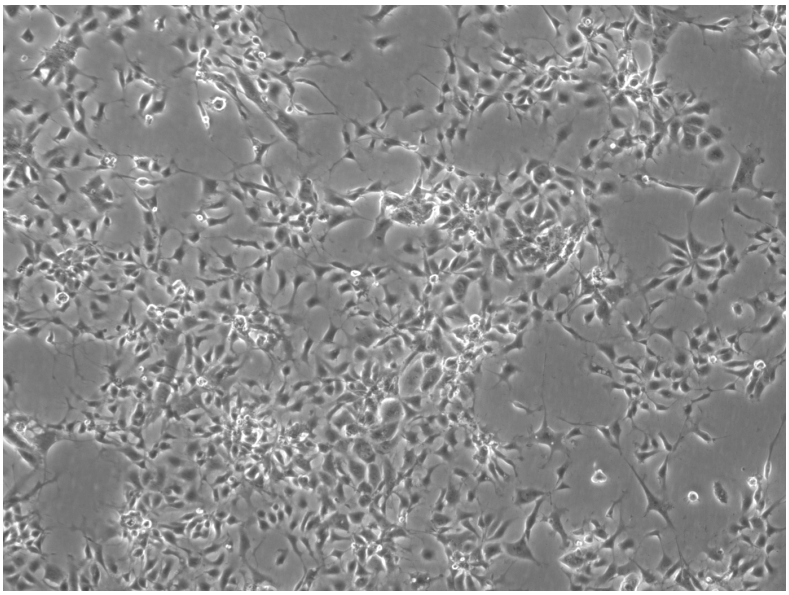


Figure 2.15. Light micrograph of 109Q neural stem cells.
10x magnification.

2.2.4 U20S cells

U20S cells were cultured in DMEM GlutaMAX (Thermo, cat #10565018) supplemented with 10% fetal bovine serum (FBS), 100 U/ml penicillin and 100 µg/ml streptomycin.

2.3 Cell imaging

2.3.1 Fixation

1-24h before fixation, 1:100 Geltrex was added to media to aid attachment. 75% of media was removed, 150 µL prewarmed 4% formalin was added, then 100 µL of this solution was replaced with fresh formalin. Cells were incubated at room temperature for 15 min, then 75% of the formalin was removed and the cells gently washed in PBS, before adding 100 µL PBS, containing 0.02% (w/v) sodium azide.

2.3.2 Immunofluoresence

Cells were permeabilised in 0.2% triton X100 diluted in PBS for 15 min at room temperature. They were incubated in blocking buffer containing 10% goat serum and 1% bovine serum albumin (BSA) diluted in PBS for 1 hour at room temperature. Primary antibodies were diluted in 1% BSA and added to cells, either for 2 hours at room temperature or overnight at 4°C. After five 5 min washes in PBS, secondary antibodies were added, diluted 1:1000 in PBS, for 2 hours at room temperature and protected from light. After three 5 min washes, cells were counterstained with Hoechst (Thermo, cat #33342), diluted 1:2000 in PBS, for 10 min. After three 5 min washes they were mounted in Dako fluorescence mounting medium (Agilent, cat #S3023). They were imaged on a Zeiss LSM 710 confocal microscope and analysed with Zen software.

2.4 Genetics

2.4.1 DNA extraction

DNA was extracted using the QIAamp DNA mini kit (Qiagen, cat #51306), according to manufacturer instructions. Briefly cell pellets were resuspended in 200 µL PBS, then 20 µL proteinase K and 200 µL of buffer AL were added, vortexed and

incubated at 56 ˚C for 10 min. 200 µL ethanol was added and the sample was transferred to the QIAamp mini spin column for centrifugation at 6000 xg for 1min. The filter was washed with buffer AW1 then AW2, and then DNA was eluted in 50 µL of buffer AE. DNA concentration was measured by nanodrop.

2.4.2 CAG repeat sizing

The sizing assay, termed fragment analysis, involves amplification of the CAG repeat region by PCR using a fluorescently labelled forward primer located upstream of the CAG region, then sizing of the fragment by capillary electrophoresis, with sufficient resolution to separate alleles of one repeat difference.

Two methods are used; **triplet-repeat primed PCR (TP PCR)** and **CAGCCG PCR** (Losekoot et al., 2013). TP-PCR is used for accurate sizing and includes a reverse primer that binds 5' of the polyproline repeat. This chimeric reverse primer is located partially within the CAG region, and hybridises to multiple locations within the CAG repeat, creating a series of PCR products that differ in size by 1 CAG, giving a characteristic ladder on the capillary electrophoresis trace. These stutter peaks extend from the smaller allele and terminate with the larger allele. This enables the detection of large pathogenic repeats that cannot be amplified by flanking primers. Because the 5' end of the reverse chimeric primer exactly matches the sequence 3' of the CAG region, this product is preferentially amplified and the highest peak represents the true allele. Neither TP-PCR or CAGCCG PCR allows amplification of alleles over around 160 CAG, so without the TP-PCR reverse primer, traditional PCR would amplify only the normal allele and the subject would appear to be homozygous for the normal allele (Bean and Bayrak-Toydemir, 2014, Bates et al., 2014).

In CAGCCG PCR, the reverse primer binds 3' of the polyproline. This improves PCR amplification, but can lead to misclassification of alleles due to variability in the number of prolines. Most people have 7 (67%) or 10 prolines (30%), but up to 12 prolines have been observed (0.5%) (Andrew et al., 1994b).

Sizing using both TP-PCR and CAGCCG PCR allows resolution of two same-sized non-pathogenic alleles by their heterozygosity for the polyproline repeat, thereby excluding failed amplification of a large pathogenic allele.

Several variants have been described in or near the CAG repeat at a collective frequency of around 1% (Bean and Bayrak-Toydemir, 2014, Gellera et al., 1996, Margolis et al., 1999). Depending on the stringency of the PCR conditions, alleles carrying a polymorphism in the primer binding sites may not be amplified, which could mimic homozygosity.

2.4.2.1 Primers

Numerous primers have been used over the years.

6-FAM labelled forward primers

- Hu4: ATGGCGACCCTGGAAAAGCTGATGAA (Quarrell et al. (2012))
- HD1: ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC (Warner et al. (1993))
- HD344: CCTTCGAGTCCCTCAAGTCCTTC (Andrew et al. (1993))
- HD3F : CCTTCGAGTCCCTCAAGTCCTT (Teo et al. (2008))
- GBF: GAGTCCCTCAAGTCCTTCCAGCA (Gonitel et al. (2008))

TP PCR reverse primer 5' of the polyproline repeat

- HDE (HD3): GGCGGTGGCGGCTGTTGCTGCTGCTGCTGC (Warner et al., 1993)

CAGCCG reverse primers 3' of the polyproline repeat

- HD482: GGCTGAGGAAGCTGAGGAG (Andrew et al. (1993))
- HD5 Losekoot: GGCGGCGGCTGAGGAAGCTGAG (Losekoot et al. (2013))
- HD5 Mangiarini: CGGCTGAGGCAGCAGCGGCTGT (Mangiarini et al. (1996))
- GBR: GCCCAAACTCACGGTCGGT (Gonitel et al. (2008))

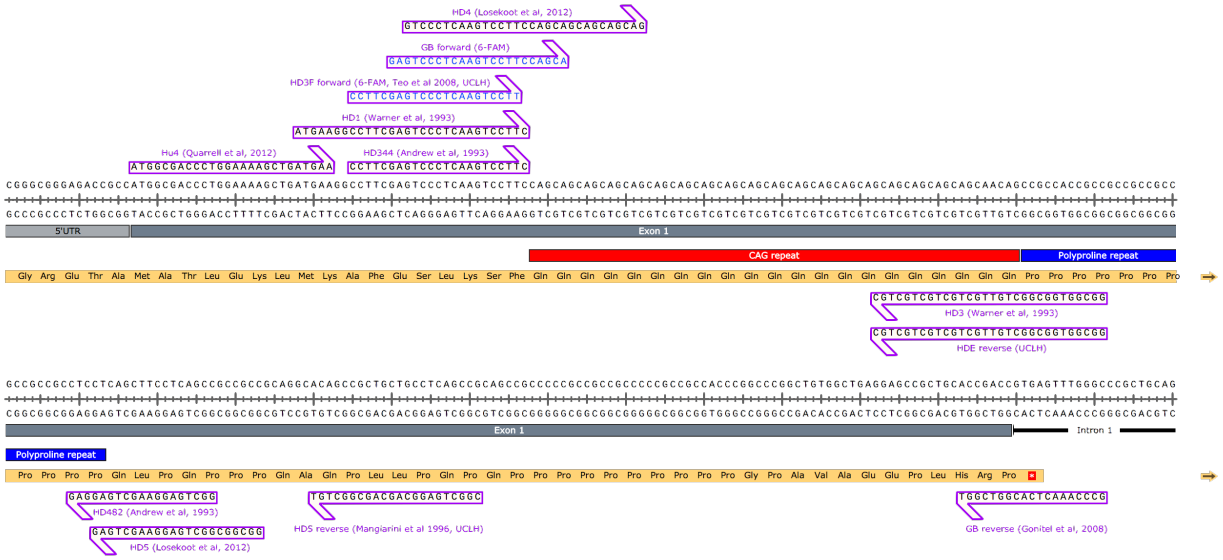


Figure 2.16. Primers for CAG repeat sizing.
The sense, antisense and amino acid sequences are given for primers targeting the CAG repeat region in HTT exon 1. Primer *are in purple, the polyglutamine tract is marked in red and the polyproline repeat in blue.*

For TP-PCR, optimal amplification was achieved the **HD3F/HDE** primer pair, and for CAGCCG PCR the **HD3F/HD5 Mangiarini** primer pair were best. These pairs were used throughout this thesis.

2.4.2.2 PCR

2.4.2.2.1 Tabrizi lab protocol

25 µL reactions were prepared containing 12.5 µL AmpliTaq Gold 360 Master Mix (Cat #4398876), 2.5 µL GC enhancer (Cat #4398848), 1 µL of each forward and reverse primer (stock 5 µM), 1µL of 50 ng/µL DNA and 7 µL water. The following cycling conditions were used.

1. Initial denaturation 95°C 10 min
2. Denature 95°C 30 sec
3. Anneal 58°C 30 sec
4. Extend 72°C 30 sec (90 sec for alleles >100Q)
5. Cycle to step 2 total 30x
6. Final extension 72°C 7 min
7. Store 4°C

2.4.2.2.2 Bates lab protocol

10 µL reactions were prepared containing 0.1 µL AmpliTaq (stock 5 U/µl), 1 µL of 2 mM dNTPs, 0.8 µL of each primer (stock 10 µM), 1 µL DMSO, 2mM AM buffer (including 0.0035 µL β-mercaptoethanol freshly added) and 2 µL of 50-100 ng/µL DNA. The following cycling conditions were used.

1. Initial denaturation 94°C 90 sec
2. Denature 94°C 30 sec
3. Anneal 65°C 30 sec
4. Extend 72°C 90 sec
5. Cycle to step 2 total 35x
6. Final extension 72°C 10 min
7. Store 4°C

2.4.2.3 Capillary electrophoresis

1 µL of PCR product was added to 10 µL Hi-Di formamide (Thermo #4311320) and the appropriate volume of size standard in a low profile non-skirted 96 well plate. For alleles <130 CAG repeats, 0.5 µL of the GeneScan™ 500 LIZ™ (Thermo #4322682) size standard was used. For alleles >130 CAG, either 0.1 µL of MapMarker ROX 1000 (Eurogentec #MW-0195-80ROX), or 0.5 µL GeneScan™ 1200 LIZ™ (Thermo #4379950) was used. The sample was denatured at 95°C for 5 min, before cooling at 4°C for 5 min. The PCR products were resolved by capillary electrophoresis on an ABI Genetic Analyzer.

2.4.2.4 Calculation of repeat size

The resulting .fsa files were displayed in GeneMapper and appear as a cluster of peaks differing by one CAG unit due to PCR stutter and mosaicism within the tissue. Allele size is assigned to the highest (modal) peak in the cluster (Swami et al., 2009). Because the number of bases flanking the CAG repeat is known, it is straightforward to calculate the number of pure CAG repeats from the PCR product size. The polyglutamine tract in most cases terminates with CAACAG, meaning there are more glutamines than CAG repeats. By convention, the number of uninterrupted CAG repeats is used as the result of the genetic test because this has a larger effect on phenotype than the number of glutamines in the repeat tract (Losekoot et al., 2013, Bean and Bayrak-Toydemir, 2014, Bates et al., 2015c). The UCLH neurogenetics lab participate in quality control assessment and have adapted the calculation, below, based on the use of international standards.

Assay	Forward primer	Reverse primer	Bases flanking pure CAG	Calculation of pure CAG repeat length	UCLH neurogenetics lab quality-controlled protocol
TP-PCR	HD3F	HDE	40	((product – 40) / 3)	((product – 39) / 3) + 2
CAG CCG	HD3F	HD5 Mangiarini	110	((product – 110) / 3)	((product – 108) / 3) + 2

Table 2.3. Calculation of CAG repeat size.

2.4.2.5 Error limits

American College of Medical Genetics (ACMG) define the measurement limits as ±2 for <50 CAG repeats, ±3 for 50-75, and ±4 for >75 (Bean and Bayrak-Toydemir, 2014). European Molecular Genetic Quality Network (EMQN) guidance advise ±1 for ≤42 CAG and ±3 for alleles >42 CAG (Losekoot et al., 2013).

The same DNA sample was run 12 times for induced pluripotent stem cells (iPSC) with 73 or 109 CAG repeats, ReNeuron VM neural stem cells with 129 CAG repeats, and patient-derived lymphoblastoid (LB) cells with 130 CAG repeats. Note, repeat instability means CAG length in these lines at the time of testing was 73, 121, 131 and 136 CAG repeats respectively. Error margins were comfortably within ACMG and EMQN guidance.

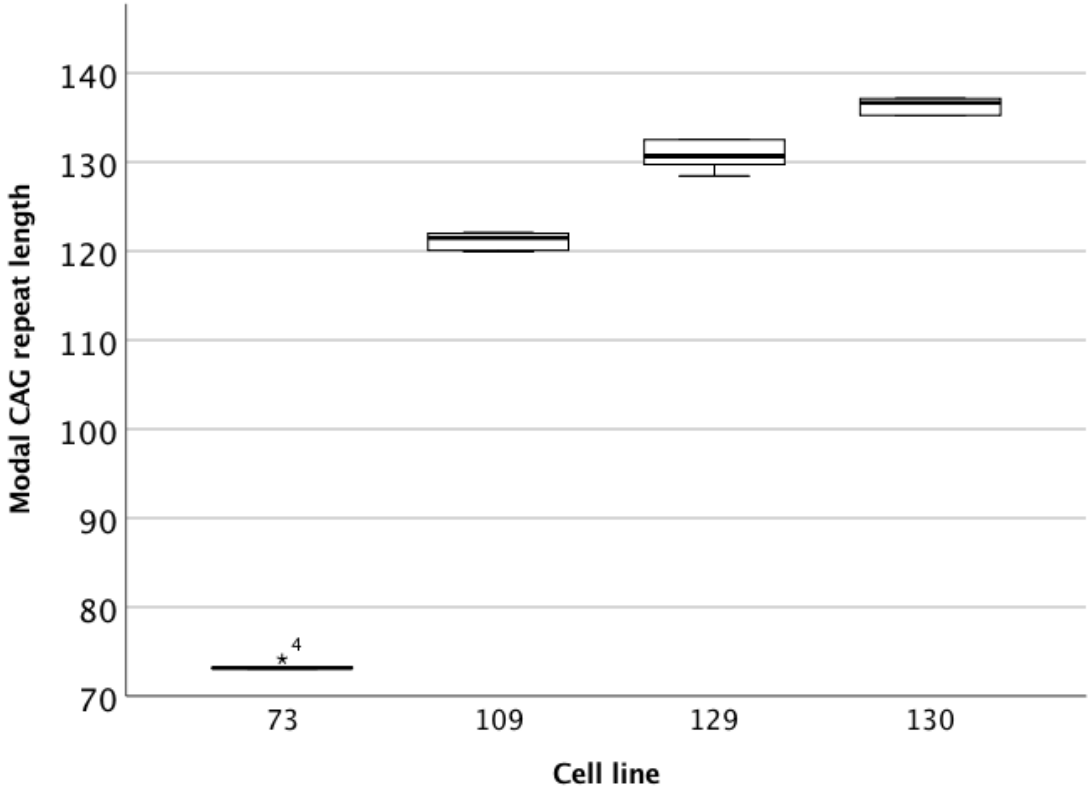


Figure 2.17. Boxplot of variability in CAG sizing from cell lines with a range of repeat lengths.
73 – 73Q iPSC, 109 – 109Q iPSC, 129 – ReNeuron VM 129Q, 130 – 130Q LB. Note, repeat instability re CAG length in these cell lines at the time of testing was 73, 121, 131 and 136 CAG repeats respectively.

	iPSC 73Q	iPSC 109Q	ReN VM 129Q	LB 130Q
Mean CAG repeat length	73	121	131	136
SD	0.4	1.0	1.5	0.9
SEM	0.17	0.40	0.51	0.38

Table 2.4. Variability in CAG sizing from cell lines with a range of repeat lengths.
Note, repeat instability means CAG length in these cell lines at the time of testing was 73, 121, 131 and 136 CAG repeats respectively.

2.4.2.6 CAG expansion analysis

2.4.2.6.1 Change in modal CAG repeat length

The primary measure of repeat instability was change in modal CAG repeat number compared to the baseline or control sample. Secondary measures included the somatic instability index (SII) (Mollersen et al., 2010) and a proportional expansion analysis. Modal CAG repeat size, SII and the proportional expansion analysis were calculated using a custom R script, available at http://caginstability.ml:3838/app/, and confirmed manually.

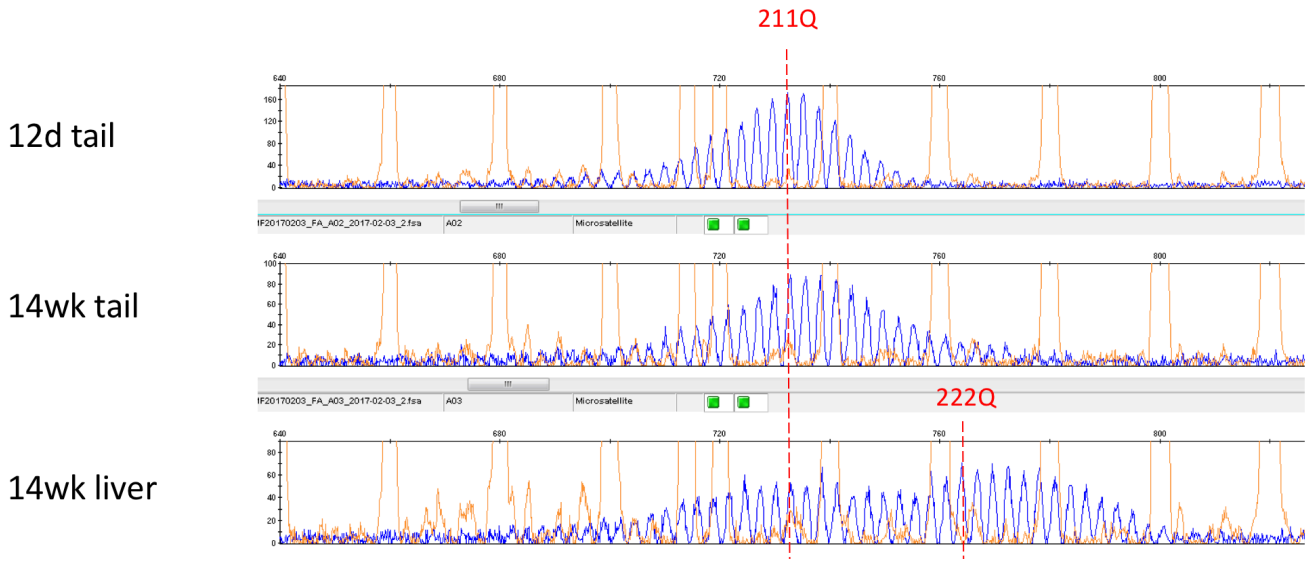


Figure 2.18. Change in modal CAG repeat length.
CAG repeat PCR generates a normal distribution of amplicon fragments spaced three base pairs apart due to PCR stutter and mosaicism within the sample tissue. Repeat size is given as the modal peak height. Change in mode is difference between baseline mode (upper trace, 12 day tail) and the mode in the experimental sample (e.g. lower trace, 14 week liver). In the example above the change in mode is 11 repeats.

2.4.2.6.2 Somatic instability index (SII)

The SII calculates expansion relative to a control sample and is measured in CAG repeat units. Firstly, the modal peak in the baseline sample is identified. Peaks less than 20% this height are excluded to remove background signal. The normalised height of each peak is calculated as a proportion of the sum of the peak heights. The change in CAG length of each peak is calculated relative to the modal peak. Normalised peak height is multiplied by change in repeat length, and the sum of these values gives the somatic instability index.

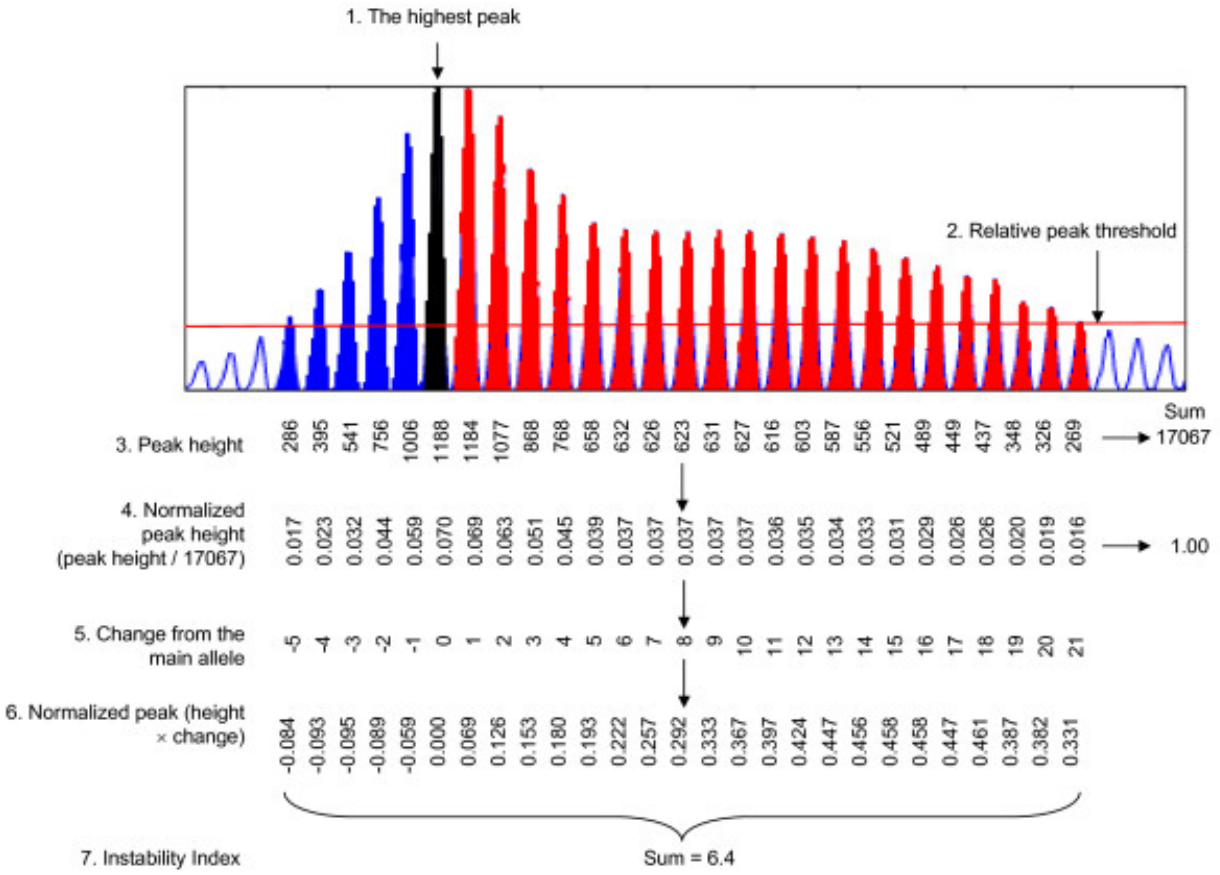


Figure 2.19. Instability index calculation from Lee et al. (2010).

Instability was quantified from GeneMapper traces. Threshold was set at 20% of the modal peak height. Peaks falling below this were excluded from analysis. Peak heights normalised to the total of all peak heights were multiplied by the change in CAG length of each peak relative to the modal peak in tail. These values were summed to generate an instability index. Striatum analysis is shown as an example (HdhQ111/+, 5 months). Open, blue, black, and red peaks represent background, contracted alleles, modal allele from tail analysis of same mouse, and expanded alleles, respectively.

2.4.2.6.3 Proportional expansion analysis

For the proportional expansion analysis, A is taken as the width of the distribution below the modal peak in the baseline tissue, and B as the width above the baseline modal peak. The modal point of the control sample is fixed, and A and B recalculated for each tissue or sample. B/(A+B) gives the proportion of the distribution that has expanded relative to the baseline sample.

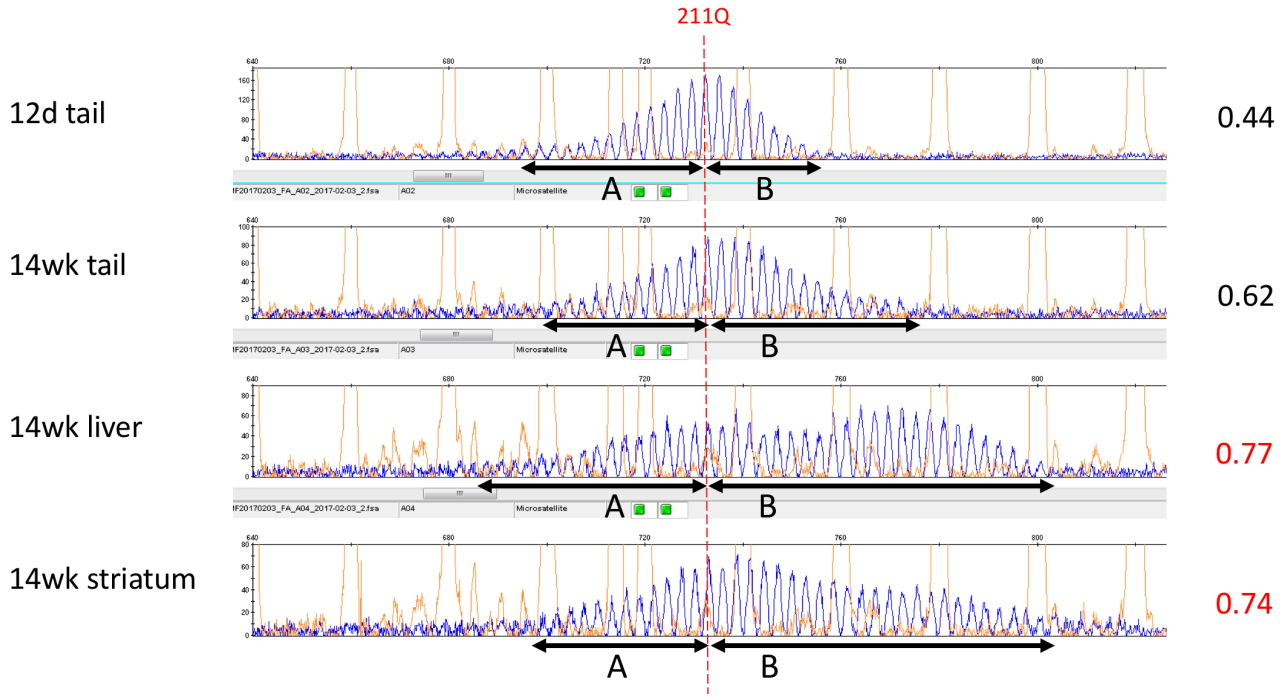


Figure 2.20. Proportional expansion analysis.

A is taken as the width of the distribution below the modal peak in the baseline tissue, and B as the proportion above the baseline modal peak. The modal point is fixed, and A and B recalculated using each tissue. B/(A+B) gives the proportion of the distribution that has expanded relative to the baseline sample.

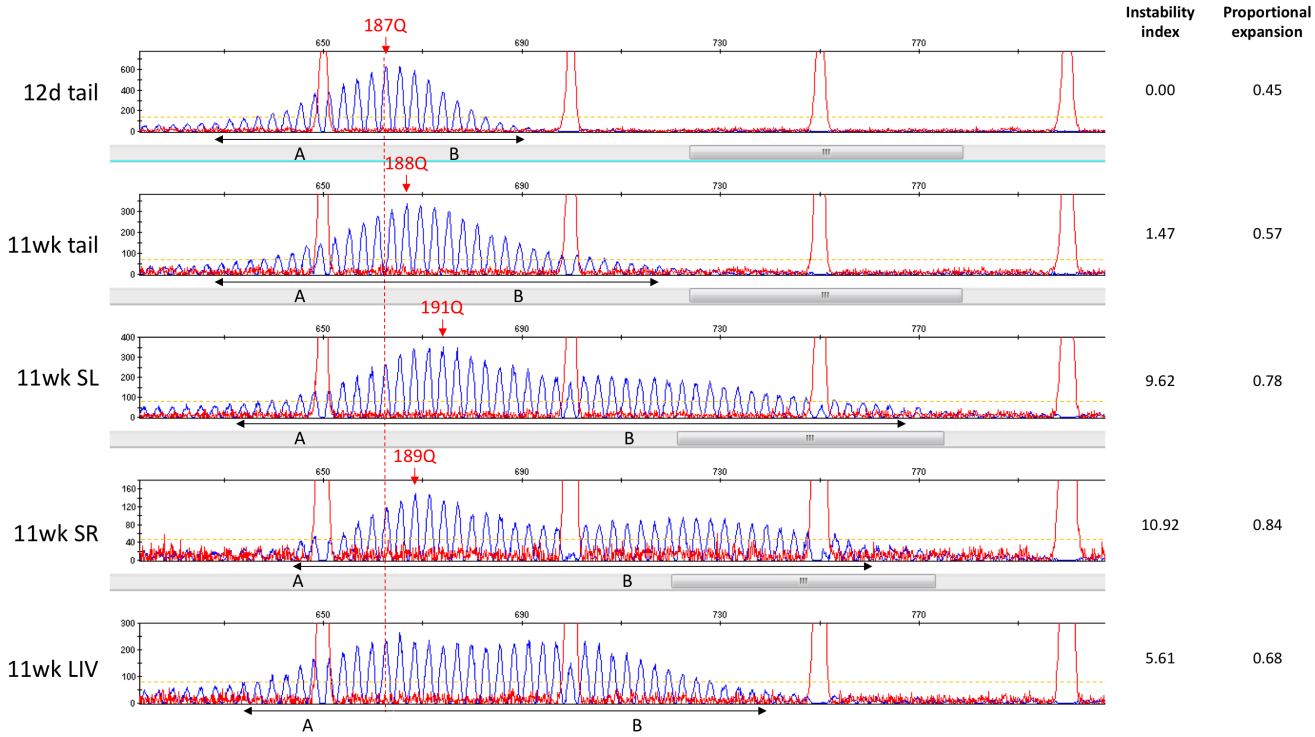


Figure 2.21. Representative example of fragment analysis traces from mouse #79 for the tissues and ages indicated (left).

Modal CAG repeat length is given by red arrows. 20% threshold value for instability index is shown by yellow hashed line. Distribution widths for proportional expansion analysis are given below each trace in black. Instability index and proportional expansion measures are shown to the right. SL – left striatum, SR – right striatum, LIV – liver. Plots displayed in GeneMapper.

2.4.2.7 Statistical analysis

2.4.2.7.1 Curve estimation

Curve estimation was performed in R by producing regression statistics for different models with a custom script (see Appendix). Variance from the model and significance was measured by analysis-of-variance.

2.4.2.7.2 Expansion rate

Expansion rate was calculated from the slope of linear regression model fitted to change in modal CAG repeat length, and was expressed as number of days per change in modal CAG length (days/Q) +/- 95% confidence interval. Expansion rates were compared with control cells by analysis of variance between the slopes of the linear regression lines in SPSS (IBM). To compare expansion rates between treatments, the rate of CAG expansion was modelled as a linear regression of CAG expansion on the number of days since the start of treatment. Differences in rate of CAG expansion between treatments were modelled by the inclusion of a day*treatment interaction term in the regression, and the significance of this term determines the significance of CAG expansion rate differences between treatments. Correlations between multiple CAG expansion measurements from the same replicate cell line were modelled by including cell line specific random effects on CAG expansion rate in the analysis. Models were fitted using the lmer function in R. Initially, global differences in CAG expansion rate were tested between all treatments simultaneously. If this test was significant, post-hoc tests were performed to characterise the differences.

2.4.3 Sanger sequencing

The target region was PCR amplified by adding 1 µL of template DNA at 200 pg/µL to 12.5 µL FastStart PCR Master mix (Sigma, cat #4710436001), 1 µL of each primer at 10 µM and 9.5 µL of water. Thermal cycling conditions were 95°C for 4 min, then 30 touchdown cycles of 95°C for 30 sec, 62°C for 30 sec and reducing by 0.2°C each cycle (final 57°C), and 72°C for 2 min. Primers and nucleotides were degraded by adding 0.02 µL exonuclease I (Thermo, cat #EN0581), 0.08 µL FastAP (Thermo, cat #EF0654) and 0.3 µL water to each 1 µL of PCR product, then incubating at 37°C for 30 min and 80°C for 15 min. The sequencing reaction was set up in a low profile, non-skirted 96 well plate by adding 1 µL of PCR product at 20 ng/µL to 1 µL of BigDye (Thermo, cat #4337455), 5 µL Better buffer (Clent, cat #3BB-5), 0.4 µL of forward or reverse primer at 10 µM, 3.8 µL of betaine or Q solution and 3.8 µL of water. Thermal cycling conditions were 96°C for 1 min, then 25 cycles of 96°C for 10 sec, 50°C for 5 sec and 60°C for 4 min. DNA was precipitated by adding 3.75 µL of 0.125 M EDTA and 45 µL of ethanol, mixing, then incubating for 15 min at room temperature before centrifuging at 3000 xg for 30 min at 4°C. The plate was then inverted and spun up to 185 xg for 30 sec. 60 µL of 70% ethanol was added and the plate was again centrifuged at 1650 xg for 15 min at 4°C before inverting it and spinning at 185 xg for 1 min. The plate was then covered to protect from light and left to dry at room temperature for 15 min or on a heat block at 37°C for 5 min. 10 µL of Hi-Di formamide (Thermo, cat #4311320) was added to each well and at least 10 µL of water to any empty wells, then the plate was sealed, vortexed and pulse centrifuged ahead of incubation at 95°C for 2 min, then cooled to 4°C for 2 min. Samples were sequenced on an ABI 3730 DNA analyzer.

2.4.4 RNA extraction

500 µL of TRIzol (ThermoFisher) was added to cells, which were triturated several times to homogenise, then incubated for 5 min. 100 µL chloroform (VWR) was added, and incubated for 5 min. The sample was centrifuged at 13,000 rpm for 15 min. The aqueous phase was transferred to a new tube and an equal volume of 70% ethanol added. RNA was purified

as per the RNeasy Mini Kit protocol (QIAGEN, 74106). A 15 min genomic DNA digestion step (DNAse I, QIAGEN, 79254) was performed between the RW1 buffer washes. RNA was eluted with water and concentration was measured on a NanoDrop 1000.

2.4.5 Quantative real time PCR (qPCR)

RNA was reverse transcribed with an initial 12 µL reaction containing 10 µL of RNA at 1 µg/µL, 1 µL of 10 mM dNTPs (ThermoFisher), and 1 µL of 100 ng/µL random hexamers P21 (Eurofins), which was incubated at 65°C for 5 min. For the second 20 µL reaction 4µL of 5x first strand buffer, 2 µL of 100mM DTT (ThermoFisher), 1µL of RNasin (Promega) and 1 µL of MMLV-RT (ThermoFisher) were added, then incubated at 25°C for 10 min, 37°C for 50 min and 70°C for 15 min(Benn et al., 2008a). The resulting cDNA was diluted to 1:10.

2.4.5.1 Taqman

15 µL qPCR reactions were set up in triplicate containing 7.5µL TaqMan Fast Advanced Master Mix (ThermoFisher), 3.75 µL water, 3 µL cDNA and 0.75 µL of the appropriate probe set (ThermoFisher). PCR was quantified on the CFX 96 qPCR system (Bio-Rad) using the following thermal cycling conditions; 95°C for 40 sec, then 40 cycles of 95°C for 7 sec and 60°C for 20 sec.

2.4.5.2 SYBR green

25 µL reactions were set up in triplicate containing 12.5 µL SYBR green master mix (Thermo, cat #4309155), 0.75 µL of each primer at 5 µM, and 10 µL water. Standard cycling conditions were used; 94°C for 2 min, then 40 cycles of 94°C for 15 sec and 60°C for 1 min.

2.4.5.3 Comparative cycle threshold analysis

The geometric mean (geomean) of housekeeping genes was used as a reference in order to determine the relative expression ratio of the genes of interest using the comparative Ct method (2-ΔΔCT) (Benn et al., 2008a). Briefly, expression level was corrected for the geomean of housekeeping genes (ΔCt), then expressed relative to the control or lowest expressing sample (ΔΔCt), and finally the fold change in expression was given by 2^-ΔΔCt. Data were analysed using a student's t-test.

2.4.6 DNA repair assays

2.4.6.1 Interstrand crosslink repair

ICL repair can be assayed by treating cultured cells with DNA crosslinking agents and assaying sensitivity (MacKay et al., 2010b, Liu et al., 2010b).

Mitomycin C (MMC) is an aziridine containing chemotherapeutic which potently crosslinks DNA by alkylating guanine nucleosides. Crosslinks are repaired by homologous recombination. Previous studies in HEK293 cells have used 0-60 ng/ml for 24h (MacKay et al., 2010a, Kratz et al., 2010b, Liu et al., 2010c). Cells were treated for 16-24 hours with MMC at the indicated concentration. They were then washed in PBS and diluted in medium to 200 cells per well of a 96 well plate and viability was assessed after 10 days in culture by MTT assay.

Cisplatin is a platinum containing cytotoxic. Platinum complexes react with DNA resulting in interstrand crosslinks, as well as intrastrand crosslinks and non-functional adducts. Previous studies in HEK293 cells used 0-1 µg/ml for 24h

(MacKay et al., 2010a, Kratz et al., 2010b). Cells were treated for 24 hours at the indicated concentration, then washed in PBS and cultured for 10 days before the MTT cell viability assay.

For the γ-H2Ax assay, cells were treated with 200 ng/ml MMC or 1 µg/ml cisplatin for 2 h, then washed in PBS and cultured in medium for 24-72h. γ-H2Ax foci clearance was analysed as described in MacKay et al. (2010b). The proportion of cells in each population with more than 10 γ-H2AX foci at each time point ("γ-H2AX positive") was determined.

2.4.6.2 Mismatch repair

6-thioguianine (6-TG) can be used to assay mismatch repair. It is converted to 6-methylthioguanine (6-MTG) which is incorporated into the genome in place of guanine and directs the insertion of thymine on the daughter strand, which is a mismatch and thus provokes MMR. Thymine is removed from the daughter strand but 6-MTG remains in place on the template strand. The cycle continues over and over, with unsuccessful repair attempts ultimately resulting in cell death. Cells lacking functional MMR are resistant (Karran and Attard, 2008, Swann et al., 1996). Cells were treated for 24 hours at the indicated concentration, then washed in PBS and cultured for 10 days before the MTT cell viability assay

2.4.6.3 Single and double strand break repair

Hydrogen peroxide (H_2O_2) oxidises bases and induces single and double strand breaks (Driessens et al., 2009). It was added to cells at the indicated concentration for 30 min, then cells were washed in PBS and cultured for 10 days before MTT cell viability assay.

2.4.7 Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) is used to determine whether specific proteins are associated with particular DNA regions. DNA and its associated proteins are crosslinked, and the DNA-protein complexes are then sheared into roughly 500 bp fragments by sonication. A protein-specific antibody was used to immunoprecipitate crosslinked DNA fragments which could then be quantified by qPCR or sequenced.

Cells were collected washed in PBS then cross-linked with formaldehyde (1% final concentration) for 10 min at room temperature. Excess formaldehyde was quenched by adding glycine to a final concentration of 125 mM and incubating for 5 min. Cross-linked cells were pelleted, washed in PBS and frozen at -80°C or used directly. Cells were lysed in lysis buffer consisting of 15 mM Tris-HCL (pH7.5), 0.3 M Sucrose, 60 mM KCl, 15 mM NaCl, 5mM $MgCl_2$, 0.1 mM EGTA, 0.5 mM DDT, 0.2% IGEPAL-CA, and supplemented with protease inhibitors. Nuclei were pelleted at 20,000 xg for 20 min. Supernatants were aspirated and nuclei were resuspended in sonication buffer, consisting of 50 mM Tris pH 8.0, 10 mM EDTA and 1 % SDS, supplemented with protease inhibitors. Chromatin was fragmented by 10 cycles of 30 sec sonication in a Bioruptor apparatus. Ice water was added to the sonication bath to ensure temperature was regulated during disruption. Insoluble material was cleared by centrifugation at 20,000 xg for 10 min. This sonicated fraction was used diluted 10-fold with dilution buffer, consisting of 16.7 mM Tris pH 8.0, 1.2 mM EDTA, 167 mM NaCl and 1% Triton X-100, supplemented with protease inhibitors. This was used as the ChIP input. Overnight Immunoprecipitation at 4°C used GFP-Trap beads (ChromoTech) to capture GFP-FAN1 forms or an affinity purified FAN1 sheep polyclonal antibody (S420C) and protein G magnetic beads (20 µl of either bead per reaction). The isolated ChIP fractions were washed twice in Wash Buffer 1, consisting of 20 mM Tris-HCl pH 8.1, 50mM NaCl, 2mM EDTA, and 1% TX-100, 0.1% SDS, then once in Wash Buffer 2, consisting of 10 mM Tris.Cl pH 8.1, 150 mM NaCl, 1 mM EDTA, 1% NP40 and 1% Na deoxycholate, and then once

in Tris-EDTA. Bound material was eluted at 65°C for 30 min in 500 µL of 1% SDS and 500 µL of 0.1 M $NaHCO_3$. Cross-links were reversed in ChIP and input fractions by adding NaCl to a final concentration of 250 mM and proteinase K (2 µl of 0.5 mg/ml PK per reaction), then heating to 65°C for at least 4 h. DNA was purified using columns (Qiagen, cat #51306) and subjected to PCR using primers detailed below. DNA was quantified by SYBR Green qPCR with the indicated primers and a QuantStudio 5 real time qPCR machine.

2.4.7.1 HTT primers

Primer	Sequence	Start	End
g.5029F	CCGCTCAGGTTCTGCTTTTA	29	48
g.5057F	CCAGAGCCCCATTCATTG	57	74
g.5174R	GCCTTCATCAGCTTTTCCAG	155	174
HD3F forward (6-FAM, Teo et al 2008, UCLH)	CCTTCGAGTCCCTCAAGTCCTT	174	195
GB forward (6-FAM)	GAGTCCCTCAAGTCCTTCCAGCA	179	201
CAG repeat		197	253
HDE reverse (UCLH)	GGCGGTGGCGGCTGTTGCTGCTGCTGCTGC	241	270
g.5310R	CTGAGGAAGCTGAGGAGGC	286	304
GB reverse (Gonitel et al, 2008)	GCCCAAACTCACGGTCGGT	402	420
g.5675F	ATTCACCGAGGGGAGTCAC	669	687
g.5773R	CCCTGGTTTCTCGCAAATAA	748	767
HTT con 1F	TTTGCCCAGGGAATCTTTGC	25715	25734
HTT con 1R	TTGCAAGCGGAGAGAGAAGA	25847	25866
HTT con 2F	TGCCTTTCGAAGTTGATGCA	137819	137838
HTT con 2R	TGCCACCACGAATTTCACAA	138001	138020

Table 2.5. HTT PCR primers.

Start and end positions are relative to the genomic sequence, numbered from the start of the 5'UTR (Homo sapiens chromosome 4, GRCh38.p12 Primary Assembly, NCBI Reference Sequence: NC_000004.12). The location of the CAG repeat region is given for reference in red.

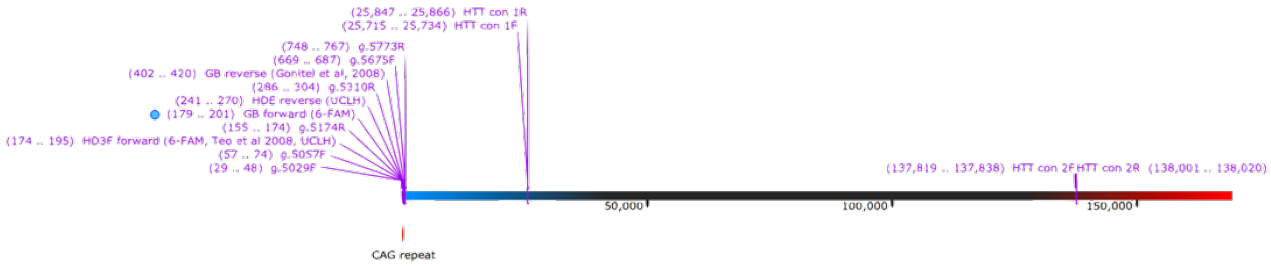


Figure 2.22. Schematic representation of HTT primers on the genomic sequence.

Figure 2.23. HTT CAG repeat primers marked on the genomic sequence.
The polyglutamine repeat is given in red.

Primer	Sequence	Start	End
ATXN3 con F	TATCCGTCTTGCAAGGTGGT	7359	7378
ATXN3 con R	CCCTGAATTGACGGCAGATG	7515	7534
ATXN3 CAG F	TTCAGACAGCAGCAAAAGCA	35564	35583
CAG repeat		35582	35611
ATXN3 CAG R	AAAGTGTGAAGGTAGCGAACAT	35711	35732

Table 2.6. ATXN3 PCR primers.
Start and end positions are relative to the genomic sequence, numbered from the start of the 5'UTR (Homo sapiens chromosome 4, GRCh38.p12 Primary Assembly, NCBI Reference Sequence: NC_000014.9). The location of the CAG repeat region is given for reference in red.

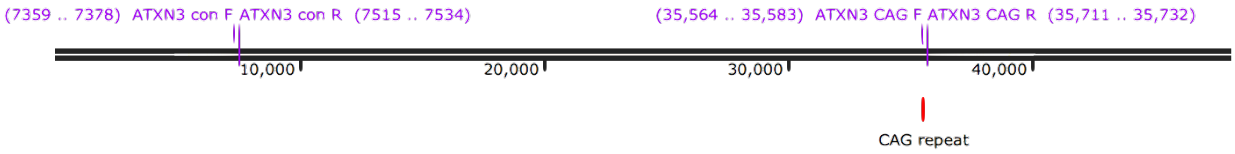


Figure 2.24. Schematic representation of ATXN3 primers on the genomic sequence.

Primer	Sequence	Start	End
DMPK con F	TGGGCCCAAAGACTCCTAAG	7722	7741
DMPK con R	TCTGAAGTCCTGTGGCTCTG	7874	7893
CTG repeat		12294	12353

Table 2.7. DMPK PCR primers.
Start and end positions are relative to the genomic sequence, numbered from the start of the 5'UTR (Homo sapiens chromosome 4, GRCh38.p12 Primary Assembly, NCBI Reference Sequence: NC_000019.10). The location of the CTG repeat region is given for reference in red.

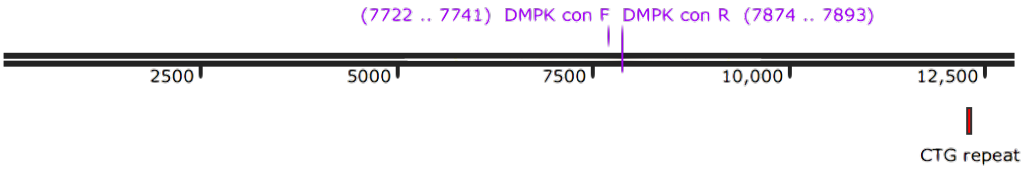


Figure 2.25. Schematic representation of DMPK primers on the genomic sequence.

Primer	Sequence	Start	End
FXN GAA F	CACTTTGGGAGGCCTAGGAA	1614	1633
GAA repeat		1725	1742
FXN GAA R	CGCCCGGCTAACTTTTCTTT	1746	1765
FXN con F	AAGCGTGCATTTTGGATTCAA	19300	19320
FXN con R	TTTTCAATTCCCTCACTGTCCTT	19466	19488

Table 2.8. FXN PCR primers.

Start and end positions are relative to the genomic sequence, numbered from the start of the 5'UTR (Homo sapiens chromosome 4, GRCh38.p12 Primary Assembly, NCBI Reference Sequence: NC_000009.12). The location of the GAA repeat region is given for reference in red.

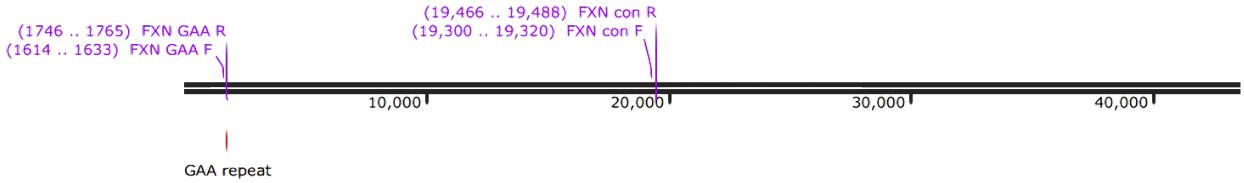


Figure 2.26. Schematic representation of FXN primers on the genomic sequence.

Primer	Sequence	Start	End
TBP con F	AAGAGTGTGCTGGAGATGCT	2605	2624
TBP con R	ATGCCCTTCCTTGCCTTTTG	2812	2831
TBP CAG F1	CAGCCAGCCTAACCTGTTTT	7421	7440
TBP CAG F2	TGACCCCACAGCCTATTCAG	7517	7536
TBP CAG R1	CTGCCTTTGTTGCTCTTCCA	7559	7578
CAG repeat		7567	7689
TBP CAG R2	TGGGACGTTGACTGCTGAA	7709	7727

Table 2.9. TBP PCR primers.

Start and end positions are relative to the genomic sequence, numbered from the start of the 5'UTR (Homo sapiens chromosome 4, GRCh38.p12 Primary Assembly, NCBI Reference Sequence: NC_000006.12). The location of the CAG repeat region is given for reference in red.

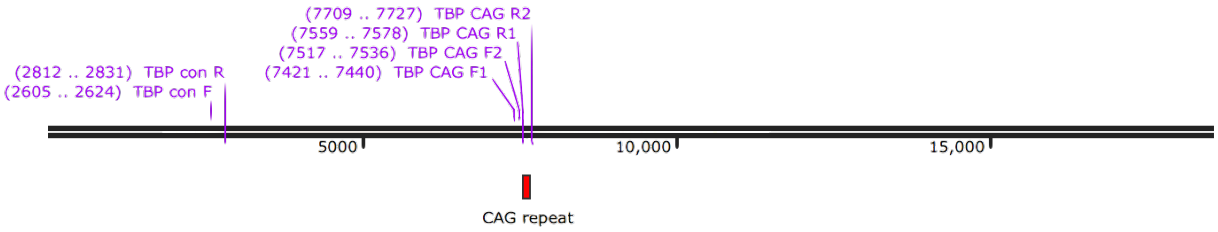


Figure 2.27. Schematic representation of TBP primers on the genomic sequence.

2.4.7.2 qPCR analysis

ChIP-qPCR data requires normalisation for sources of variability, including the amount of chromatin loaded, the efficiency of immunoprecipitation and DNA recovery. Data was normalised using the fold enrichment method (Scientific, 2018, Lacazette, 2017). The control immunoprecipitation (IP) without antibody signal is subtracted from the immunoprecipitation sample (dCt), then the fold enrichment is calculated as 2^{-dCt}. This method assumes the background signal is reproducible between samples.

2.5 Protein

2.5.1 Western blotting

Cells were washed in PBS before centrifugation at 13,000 G for 1 min and removal of supernatant. The pellet was resuspended in 200 µL of ice-cold lysis buffer, consisting of 0.1% Triton X-100, 0.1% sodium deoxycholate, 50 units/ml of benzonase, protease inhibitors, and incubated on ice for 20 min, agitating regularly. A small volume was removed for protein quantification assay using Bio-Rad assay. Proteins were precipitated overnight in 800 µL cold methanol, then centrifuged for 20 min at 12,000 rpm and 4°C and the supernatant removed. Pellets were dried, then resuspended in at a uniform concentration in SDS sample buffer (50 mM Tris-HCl pH 6.8, 2% SDS, 10% glycerol, 1% β-mercaptoethanol, 12.5 mM EDTA and 0.02% bromophenol blue), shaking at 95°C for 10 min to assist resuspension. Equal amounts of protein were loaded into the wells of a 9% SDS-polyacrylamide stacking gel (9% gel: 2.5 ml gel loading buffer, 6 ml of 30% acrylamide, 0.1% SDS, 10 µL of tetramethylethylenediamine (TEMED), 100 µL ammonium persulfate (APS), made up to 10 ml with water, 1 mm plates. Stacking gel: 312.5 µL acrylamide, 312.5 µL stacking gel buffer, 0.1% SDS, 2.5 µL TEMED, 125 µL APS, made up to 2.5 ml with water). The gel was run at 60 V for 10 min, until samples had run through the separating gel, then at 125 V for 1-2h in running buffer (25 mM Tris, 190 mM glycine, 0.1% SDS, pH 8.3). Samples were then transferred to nitrocellulose membrane in transfer buffer (25 mM Tris base, 190 mM glycine, 20% methanol, pH 8.3). The membrane was blocked for 1h at room temperature or overnight at 4°C in Odyssey blocking buffer (Li-cor, #927-40100). The appropriate concentration of primary antibody was added in blocking buffer and incubated overnight at 4°C. The membrane was washed five times in TBS-T (1x TBS, 900 ml water, 0.1% Tween 20), 5 min each wash, then incubated with secondary antibody in blocking buffer at room temperature for 1h. The membrane was then again washed three times in TBS-T (Goold et al., 2011), and imaged using an Odyssey Infrared Imager (LI-COR).

Chapter 3 DNA repair variants modify phenotype in polyglutamine diseases

3.1 Background

3.1.1 Polyglutamine diseases

The 9 polyglutamine diseases are HD, SCA types 1, 2, 3 (Machado-Joseph disease, MJD), 6, 7 and 17, dentatorubral pallidoluysian atrophy (DRPLA), and spinal bulbar muscular atrophy X-linked type 1 (SMAX1, or SBMA). HD and SCA3 are the commonest, but the mean prevalence of all polyglutamine diseases is around 1/10,000 (Fan et al., 2014). All are dominantly inherited, with the exception of SBMA, which is X-linked. The protein in each case contains an expanded glutamine tract which leads to aggregation. In all, there is a negative correlation between repeat length and disease severity, all show anticipation, with the repeat tending to expand and cause more severe disease intergenerationally, all tend to begin in middle life, progress inexorably to death within around 15-20 years and have a predominantly neurological phenotype (Fan et al., 2014).

Polyglutamine disease	Gene	Prevalence (per 100,0000)	Variance in AAO explained by repeat length (% that is heritable)	Normal range repeat length	Pathogenic range	Somatic instability
HD	*HTT*	3–10 (Wood, 2012, Rawlins, 2010, Bates et al., 2014, Warby et al., 2014)	50–60% (Gusella et al., 2014, Persichetti et al., 1994, Snell et al., 1993, Andrew et al., 1993, Duyao et al., 1993) (40–60% (Djousse et al., 2003, Wexler et al., 2004b))	6–35	40–121	Yes
SCA1	*ATXN1*	0.16 (Durr, 2010)	64–76% (Ranum et al., 1994, Tezenas du Montcel et al., 2014, Globas et al., 2008, van de Warrenburg et al., 2002, van de Warrenburg et al., 2005) (no detected heritable component (van de Warrenburg et al., 2005))	6–38	45–83	Yes
SCA2	*ATXN2*	0.2 (Durr, 2010)	50–80% (Giunti et al., 1998, Tezenas du Montcel et al., 2014, Velazquez Perez et al., 2009, Globas et al., 2008, Hayes et al., 2000, Geschwind et al., 1997, Pulst et al., 2005, van de Warrenburg et al., 2002, van de Warrenburg et al., 2005, Lorenzetti et al., 1997) (17–59% (Pulst et al., 2005, van de Warrenburg et al., 2005))	15–31	33–500	Yes
SCA3	*ATXN3*	0.4 (Durr, 2010)	45–80% (Tezenas du Montcel et al., 2014, Maruyama et al., 1995, Bettencourt and Lima, 2011, Globas et al., 2008, Riess et al., 2008, van de Warrenburg et al., 2002, Durr et al., 1996, van de Warrenburg et al., 2005) (46% (van de Warrenburg et al., 2005))	12–44	52–87	Yes
SCA6	*CACNA1A*	0.04 (Durr, 2010)	26–52% (Tezenas du Montcel et al., 2014, Globas et al., 2008, Matsuyama et al., 1997, van de Warrenburg et al., 2002, van de Warrenburg et al., 2005) (no detected heritable component (van de Warrenburg et al., 2005))	4–18	20–33	Unknown
SCA7	*ATXN7*	0.12 (Durr, 2010)	71–88% (Tezenas du Montcel et al., 2014, David et al., 1998, van de Warrenburg et al., 2002, van de Warrenburg et al., 2005) (no detected heritable component (van de Warrenburg et al., 2005))	3–19	37–460	Yes
SCA17	*TBP*	<0.02 (Durr, 2010)	Unknown	25–40	49–66	Unknown
DRPLA	*ATN1*	0.005–0.04 (Le Ber et al., 2003, Filla et al., 2000, Pujana et al., 1999, Silveira et al., 2002)	50–68% (Wardle et al., 2009, Potter, 1996)	6–35	48–93	Yes
SBMA	*AR*	0.65–2.00 (Udd et al., 1998, Spada, 2014)	29% (Sinnreich et al., 2004)	9–34	38–72	Yes

Table 3.1. Characteristics of the polyglutamine diseases.

Epidemiology and CAG repeat ranges of polyglutamine diseases. Prevalence is given per 100,000 European population. AAO = age at onset; HD = Huntington's disease; SCA = spinocerebellar ataxia; DRPLA = dentatorubral-pallidoluysian atrophy; SBMA = spinal and bulbar muscular atrophy.

Polyglutamine disease	Phenotype
HD	Involuntary movements including chorea, and impaired voluntary movements with incoordination and bradykinesia (Bates et al., 2015c). Cognitive impairment in attention, mental flexibility, planning, initiation, lack of awareness, disinhibition (Duff et al., 2010) and impulsivity (Papoutsi et al., 2014). Neuropsychiatric features including depression, irritability, apathy and executive dysfunction.
SCA1	Ataxia, bulbar, somatosensory an oculomotor dysfunction, pyramidal signs, visual impairment, electrophysiological abnormalities (Subramony, 2012), executive dysfunction and verbal memory impairment (Burk et al., 2003, Kawai et al., 2009).
SCA2	Ataxia, slow eye movements, peripheral neuropathy, postural and action tremor, myoclonus and hyporeflexia, dementia with impaired attention, memory, frontal executive function (Kawai et al., 2009, Burk, 1999, Storey et al., 1999, Sokolovsky et al., 2010).
SCA3 (MJD)	Ataxia, pyramidal involvement, ophthalmoplegia, peripheral neuropathy, parkinsonism in some, cognitive dysfunction including attention difficulties (Burk et al., 2003, Maruff et al., 1996).
SCA6	Ataxia, occasional extrapyramidal and sensory effects, impaired executive function and visual memory (Kawai et al., 2008, Globas et al., 2003, Sokolovsky et al., 2010).
SCA7	Ataxia, retinal degeneration, ophthalmoplegia, seizures, and dementia with impaired executive function (Sokolovsky et al., 2010).
SCA17	Ataxia, psychiatric features, extrapyramidal signs, seizures, dementia and apraxia (Schneider et al., 2006, Rolfs et al., 2003).
DRPLA	Ataxia, choreoathetosis, myoclonus, epilepsy, and dementia (Ikeuchi et al., 1995, Naito and Oyanagi, 1982, Warner et al., 1994).
SBMA	Male only. Progressive weakness, atrophy and fasciculations in limbs and bulbar musculature (Atsuta et al., 2006, Kennedy et al., 1968, Rhodes et al., 2009), peripheral sensory neuropathy (Antonini et al., 2000), gynaecomastia, testicular atrophy and reduced fertility (Dejager et al., 2002).

Table 3.2. Phenotypes of polyglutamine diseases.
HD = Huntington's disease; SCA = spinocerebellar ataxia; DRPLA = dentatorubral-pallidoluysian atrophy; SBMA = spinal and bulbar muscular atrophy.

3.1.2 Genetic modifiers

A way of overcoming the complexity of the pathogenesis of repeat expansion disease is an unbiased search of the genome for loci that modify the course of the disease in a beneficial or damaging way. Drug manipulation of the identified biological pathway could have similarly beneficial effects on phenotype. In HD, the length of the CAG repeat in HTT explains around 56% of observed variation in age at onset (Gusella et al., 2014, Langbehn et al., 2004), but up to 40% of the remaining variability is heritable and due to genetic differences elsewhere in the genome (Wexler et al., 2004a).

3.1.2.1 Age at onset

Investigating these genetic modifiers, a GWAS of over HD 4000 patients identified three independent genome-wide loci significantly associated with age at motor onset; one on chromosome 8 and two close together on chromosome 15 (GeM-HD, 2015). In any GWA, the location of significant variants does not immediately identify which gene underlies the effect; genotyping arrays designed for GWA studies use linkage disequilibrium to provide coverage of the entire genome by genotyping a subset of variants. They identify an association signal, or a region of linkage disequilibrium containing causal variants, which can contain hundreds of variants, spanning relatively large regions that encompass several genes (Bush and Moore, 2012, Spain and Barrett, 2015). The loci on chromosome 15 could be associated with *FAN1*, *MTMR10* or the pseudogene *HERC2P10*. These lie in a region of copy number variation (CNV) due to non-homologous recombination of flanking repeats, and both deletion or duplication of the region have been associated with intellectual disability, epilepsy, autism and schizophrenia (Ionita-Laza et al., 2014). The highest priority candidate gene is *FAN1*, a nuclease that is involved in DNA interstrand crosslink (ICL) repair and interacts with MMR proteins including MLH1. The two chromosome 15 signals were in opposing directions; the minor allele at rs146353869 was associated with 6.1 year earlier onset ($p = 4.3x10^{-20}$) and at rs2140734 with 1.4 year later onset ($p = 7.1x10^{-14}$), suggesting *FAN1* polymorphisms can be protective or

damaging in HD. These effect sizes are significant, representing up to 33% of disease duration and 30% of lifespan prior to diagnosis. The top coding SNP in *FAN1*, and third most significant overall, rs150393409 (pArg507His) in the DNA binding domain, is in strong linkage disequilibrium with the index SNP, rs146353869, and is associated with 6-year early onset ($p = 9.34x10^{-18}$). Both these SNPs were imputed and either could be driving the deleterious GWAS signal. *MTMR10* is involved in inositol-phosphate signalling.

The role of FAN1 in the Fanconi anaemia ICL repair pathway remains unclear (Ceccaldi et al., 2016b). Rather than targeting specific DNA sequences, it is structure-specific, binding branched substrates that mimic DNA repair (Ceccaldi et al., 2016b). Therefore, it is conceivable that FAN1 could target abnormal DNA structures formed by expanded repeats. Its mutation does not cause Fanconi anaemia, but loss-of-function mutations lead to the recessive renal syndrome karyomegalic interstitial nephritis (KIN) (Zhou et al., 2012) and heterozygous truncating mutations can cause hereditary colorectal cancer in a similar way to MMR mutations (Segui et al., 2015a). FAN1 may act directly at the repeat expansion to influence expansion. Alternatively, deficits in DNA repair could independently add, perhaps in an activity dependent manner (Madabhushi et al., 2015), to the selective neurodegeneration induced by the *HTT* CAG repeat expansion.

The chromosome 8 association signal, driven by rs1037699 (p = 2.7E-08), contained *RRM2B* and *UBR5*. RRM2B is a subunit of the rate-limiting enzyme in deoxyribonucleotide triphosphate (dNTP) synthesis (Pontarin et al., 2011), which is involved in DNA replication and repair (Pontarin et al., 2012), regulates mitochondrial DNA content (Bourdon et al., 2007) and reduces oxidative stress (Kuo et al., 2012). *UBR5* is a E3 ubiquitin ligase involved in proteasome-mediated protein degradation, a function implicated in the accumulation of misfolded HTT fragments (Ortega and Lucas, 2014).

A peak on chromosome 3 that did not reach significance in the GeM GWAS, but which a subsequent study with a larger cohort found genome-wide significant (Lee et al., 2017), was driven by rs144287831 (p = 2.2E-7) and centres on *MLH1*. *MLH1* is known to interact with FAN1 (MacKay et al., 2010a, Kratz et al., 2010a, Liu et al., 2010c, Smogorzewska et al., 2010b) and its inactivation in HD mice reduces somatic instability and improves disease phenotype (Pinto et al., 2013a).

The pathway analysis found a substantial enrichment for the network of DNA repair genes, particularly nucleotide excision repair (p = 1.69E-06) and mismatch repair (p = 3.25E-06), independent of genome-wide significant individual SNPs.

3.1.2.2 Progression

A more recent GWAS identified a locus on chromosome 5 that was associated with the rate of HD progression (Hensman Moss et al., 2017b). This region included *MSH3*, *DHFR* and *MTRNR2L2*, and genome-wide significant variants were shown to influence *DHFR* expression in brain and peripheral tissues, and *MSH3* expression in blood and fibroblasts. The lead single nucleotide polymorphism (SNP), rs557874766 (p = 1.58E-08), lies within a 9 bp tandem repeat sequence in exon 1 of *MSH3* (Nakajima et al., 1995) and was reported to result in the coding change Pro67Ala. The chromosome 15 locus implicated by the previous GWAS and likely underlain by *FAN1* was just below genome-wide significance (p = 3.97E-4). Once again gene set analysis strongly enriched for DNA repair pathways, particularly mismatch repair (p = 8.88E-11).

3.1.3 Summary

Investigating the mechanism of repeat instability is complex and numerous processes appear to contribute. CAG repeat expansion requires mismatch repair proteins MSH2 (Savouret et al., 2004, Wheeler et al., 2003, Manley et al., 1999),

MSH3 (van den Broek et al., 2002) and PMS2 (Savouret et al., 2003). The long-term goal is to modulate instability level and direction, aiming to produce clinical benefit. It is likely that in answering these questions we will uncover novel genetic phenomena, and probably more questions.

Genetic modifiers of Huntington's disease have recently been found, highlighting parts of the DNA damage response such as mismatch repair, base excision repair and the Fanconi anaemia pathway. This raises questions about whether these processes also modify phenotype across all trinucleotide repeat diseases and if a common pathway underlies repeat instability.

3.2 Aims

Genetic variation in DNA repair proteins has been shown to influence age at motor onset and rate of progression in Huntington's disease. The work presented in this chapter aims to determine whether DNA repair genes also modify disease course in the other polyglutamine diseases too.

3.3 Methods

3.3.1 Patient cohorts

Cohorts of subjects with all 9 polyglutamine diseases were gathered from the Neurogenetics Unit and Ataxia Centre of the National Hospital for Neurology and Neurosurgery (London, UK), TRACK-HD (Europe) (Tabrizi et al., 2013), the SPATAX network (France), the University of Athens Medical School/Eginition Hospital (Athens, Greece), the National Institute of Neurology and Neurosurgery, Manuel Velasco Suarez (Mexico), and the University of Azores (Ponta Delgada, Portugal). All polyglutamine patients willing to participate in research were enrolled regardless of CAG repeat length or age at onset (AAO). Very few DRPLA and SBMA samples were available, so these diseases were excluded from the analysis.

Given the diverse range of phenotypes in polyglutamine diseases, AAO was recorded as the onset of motor symptoms in HD and as the first progressive symptom as reported by the patient for the other conditions. AAO was available for 1,462 patients.

Cohort	**Disease**							
	HD	**SCA1**	**SCA2**	**SCA3**	**SCA6**	**SCA7**	**SCA17**	**Total**
Athens, Greece	351	0	0	0	0	0	0	351
Azores, Portugal	0	0	0	91	0	0	0	91
London, UK	0	30	66	45	69	7	1	218
Mexico	0	0	113	0	0	66	6	185
SPATAX, France	0	147	115	261	0	0	0	523
TRACK-HD, Europe	94	0	0	0	0	0	0	94
Total	**445**	**177**	**294**	**397**	**69**	**73**	**7**	**1,462**
% M	49.4	54.2	48.6[a]	52.6	60.9	56.2	85.7	51.8[a]
Mean AAO ± SD (range)	45 ± 12.1 (6–82)	37 ± 10.5 (16–65)	33 ± 12.9 (8–73)	39 ± 11.6 (9–74)	57 ± 10.5 (18–76)	35 ± 17.6 (5–84)	30 ± 13.4 (8–44)	
Mean (CAG)n length ± SD (range)	44 ± 5.0 (37–92)	48 ± 5.3 (39–66)	42 ± 4.5 (33–64)	71 ± 4.4 (50–82)	22 ± 0.9 (21–26)	48 ± 11.1 (36–100)	51 ± 6.4 (42–58)	

Table 3.3. Cohort characteristics.
HD = Huntington's disease; SCA = spinocerebellar ataxia; % M = percentage of males; AAO = age at onset; SD = standard deviation.
[a]One subject had no sex information.

3.3.2 Selection of single nucleotide polymorphisms (SNPs)

SNPs were selected from the most significant genes in DNA repair pathways that were significantly associated with AAO in the GeM GWAS (GeM-HD, 2015), as listed in their supplementary table S4. A small number of the most significant variants were selected in order to minimise the statistical limitations of multiple comparisons. These variants were not intended to comprehensively cover all DNA repair genes, but rather to have a likely influence on disease course based on their significance in the GeM GWAS (GeM-HD, 2015) or previous evidence of involvement in disease pathogenesis. *MLH1*, *PMS1*, *PMS2*, *MSH3*, *FAN1*, *RRM2B*, *UBR5* and *LIG1* have all be associated with age at onset or rate of progression in HD GWA studies (Lee et al., 2019, GeM-HD, 2015, Lee et al., 2017, Hensman Moss et al., 2017b). *MLH1* (Pinto et al., 2013a), *MLH3* (Pinto et al., 2013a), *PMS2* (Gomes-Pereira, 2004, Gomes-Pereira et al., 2014b, Pinto et al., 2013b), *MSH3* (Dragileva et al., 2009, Tome et al., 2013a), *MSH6* (Dragileva et al., 2009, van den Broek et al., 2002, Foiry et al., 2006) and *LIG1* (Lopez Castel et al., 2009, Tome et al., 2011) are mismatch repair genes that are required for somatic instability in mouse models, and increased expression of *PMS1* has been associated with later HD onset (Lee et al., 2019). *RRM2B* is involved in nucleotide synthesis, and *UBR5* is a ubiquitin ligase previously investigated for a role in HTT aggregation

(Ortega and Lucas, 2014). *ERCC3* is a helicase involved in nucleotide excision repair which was significantly associated with age at onset in GeM-HD (2015).

For each gene, the most significant SNP was selected along with a small number of proxy SNPs that were in close linkage disequilibrium ($r^2 > 0.8$) and were also associated with AAO. Where possible they were chosen to have a functional impact (1000 Genomes). In the case of *FAN1*, where there were two independent signals, the second signal was also included. This yielded 22 SNPs.

SNP ID	Chr:position (bp) (GRCh37/hg19)	Gene symbol	Functional annotation	P GeM-HD)	MAF*	Genotype call rate*	P (HWE)*
rs1800937	2:48025764	MSH6	Stop_gained	0.0043	0.074	0.973	0.84
rs4150407	2:128049631	ERCC3	Intron_variant	0.00046	0.479	0.964	0.003
rs5742933	2:190649316	PMS1	NMD_transcript_variant	0.000949	0.205	0.972	1
rs1799977	3:37053568	MLH1	Missense_variant	7.16E-07	0.28	0.966	0.354
rs6151792	5:80056961	MSH3	Intron_variant	0.000209	0.117	0.978	0.706
rs115109737	5:80102444	MSH3	Intron_variant	0.00045	0.041	0.98	0.489
rs71636247	5:80118976	MSH3	Intron_variant	0.000255	0.034	0.976	1
rs1805323	7:6026942	PMS2	Missense_variant	0.0304	0.043	0.975	0.736
rs12531179	7:6028687	PMS2	Intron_variant	0.0000384	0.169	0.971	0.925
rs3735721	8:103217695	RRM2B	3'_UTR_variant	5.68E-07	0.083	0.971	0.058
rs1037700	8:103250775	RRM2B	Intron_variant	5.03E-08	0.094	0.973	0.002
rs5893603	8:103250839	RRM2B	Frameshift_variant	4.28E-08	0.093	0.973	0.007
rs1037699	8:103250930	RRM2B	Missense_variant	2.7E-08	0.094	0.976	0.002
rs16869352	8:103306033	UBR5	Synonymous_variant	4.01E-07	0.08	0.975	0.03
rs61752302	8:103311153	UBR5	Synonymous_variant	0.00303	0.026	0.977	0.621
rs72734283	14:75495059	MLH3	Intron_variant	0.00432	0.089	0.971	0.623
rs175080	14:75513828	MLH3	Missense_variant	0.00772	0.435	0.971	0.447
rs146353869	15:31126401	FAN1	Intron_variant	4.3E-20	0.017	0.973	1
rs114136100	15:31197976	FAN1	Synonymous_variant	8.49E-16	0.019	0.976	0.423
rs150393409	15:31202961	FAN1	Missense_variant	9.34E-18	0.013	0.975	1
rs3512	15:31235005	FAN1	3'_UTR_variant	5.28E-13	0.283	0.973	1
rs20579	19:48668830	LIG1	NMD_transcript_variant	0.00665	0.134	0.942	0.732

Table 3.4. Characteristics of single nucleotide polymorphisms (SNPs) used in this study.

*SNPs were selected from the most significant genes (gene-wide p<0.1) in the "DNA repair pathway cluster" from the GeM-HD analysis (GeM-HD, 2015) (listed in Table S4 of GeM-HD). Genes annotated by the SNPs are indicated. *Refers to the current study. Chr = chromosome; MAF = minor allele frequency; HWE = Hardy–Weinberg equilibrium.*

3.3.3 Genotyping

SNP assays were designed from sense or antisense sequences and genotyping was performed using custom fluorescence-based KASP (Kompetitive allele specific PCR) assays at LGC Genomics (Hertfordshire, UK). Note, SNPs rs4150407, rs1805323, rs1037700, rs1037699, rs3512 and rs20579 were genotyped in reverse orientation, so the genotypes will also be complementary to HGVS nomenclature.

SNPs	HGVS Names	SNP to Chromosome	Seed sense sequences for KASP assay design
rs1800937	NC_000002.11:g.48025764C>T	Forward	TTGCCTGGCAGGTAGGCACAACTTA[C>T]GTAACAGATAAGAGTGAAGAAGATA
rs4150407	NC_000002.11:g.128049631T>C	Reverse	AGTACACAATGGGAAGGTGGTCCAT[A>G]GACAAGAGCCTTCACCAGAAACTGA
rs5742933	NC_000002.11:g.190649316G>C	Forward	GTAATTGCCTGCCTCGCGCTAGCAG[G>C]AAGGTAGTGTGGTGTGACTAACGGG
rs1799977	NC_000003.11:g.37053568A>G	Forward	CTCAACCGTGGACAATATTCGCTCC[A>G]TCTTTGGAAATGCTGTTAGTCGGTA
rs6151792	NC_000005.9:g.80056961C>T	Forward	TCACACAGCCATGTAAAATTAGGCC[C>T]GCAGACAATTCGAAGGAGGAGAAAA
rs115109737	NC_000005.9:g.80102444G>A	Forward	GAATCACACAAGCTTATTTGCTATA[G>A]CATTATAATAACTTTTTACATCTGT
rs71636247	NC_000005.9:g.80118976A>G	Forward	TGTATAAATATATGTGGAGAAAACC[A>G]TCTAGATAGAAGGCTTATTCCAAAA
rs1805323	NC_000007.13:g.6026942G>T	Reverse	TCCAGTCACGGACCCAGTGACCCTA[C>A]GGACAGAGCGGAGGTGGAGAAGGAC
rs12531179	NC_000007.13:g.6028687C>T	Forward	ATTTTTAGTAGAGACAGAGTTTCAC[C>T]GTGTTAGATAGTCTCGATCTCCTGA
rs3735721	NC_000008.10:g.103217695A>G	Forward	GCTGGGGCCAGCTTAGTTGTAAGAA[A>G]AACTATTATTGTATATAATTGGACA
rs1037700	NC_000008.10:g.103250775G>C	Reverse	GGCCTCAGGCCGGGGGTGAGACTTAC[C>G]CCTGCGTTTATCCGCCTCACGCTCT
rs5893603	NC_000008.10:g.103250839_103250840insG	Forward	TTGGCTGGCCCCGGGGCAGAGCAGC[->G]GAGCGGGACGCAAACCCAAAGTCAG
rs1037699	NC_000008.10:g.103250930C>T	Reverse	AGGACAGGCCTGTCCGCCCGCCCTC[G>A]CCGCAGCCTGGCTTCGTCGTTGCGA
rs16869352	NC_000008.10:g.103306033T>C	Forward	CAGCGTAAGGTAGCAATGCTTGGAA[T>C]ACACGCTTGCATTTTCCAATTGGCT
rs61752302	NC_000008.10:g.103311153C>T	Forward	ACAATTTCAATATAAAATGAGCATT[C>T]GCCTTTCGATCCTTGGATTCTACTA
rs72734283	NC_000014.8:g.75495059A>G	Forward	ATTATTTTATGATTTGACCTTGACA[A>G]CCCATCTAGCCAACTCCCATCCAGT
rs175080	NC_000014.8:g.75513828G>A	Forward	GGTCATAGGACTTTCTCTCAAACTA[G>A]GCATCTGTTGTTCTAAACAATCTTC
rs146353869	NC_000015.9:g.31126401C>A	Forward	AATGGTATGTATTAAAATGTGAATC[C>A]CAAGAGTGATGTGTCACTGTGCACT
rs114136100	NC_000015.9:g.31197976C>T	Forward	GCTGCAATGGTCCTGGTCAAACAAC[C>T]GGTCATCCTTACTACCTTCGGAGTT
rs150393409	NC_000015.9:g.31202961G>A	Forward	GCCTTTCTCAAATTGGCCAAACAGC[G>A]TTCAGTCTGCACTTGGGGCAAGAAT
rs3512	NC_000015.9:g.31235005G>C	Reverse	ACAGAGAGCGTTAAAAGTAAAGGCA[C>G]TTCCAAGAGTAACACTGCTAATGCG
rs20579	NC_000019.9:g.48668830G>A	Reverse	GCTGGACAGGAAGGGAGAATTCTGA[C>T]GCCAACATGCAGCGAAGTATCATGT

Table 3.5. Seed sense sequences for SNP KASP assay design.
Note that genotypes for SNPs in reverse orientation to chromosome given by KASP assays are complementary (reverse) to HGVS nomenclature.

3.3.4 Statistical analyses

AAO for each polyglutamine disease was corrected for repeat length to derive the residual AAO which was used as the primary phenotype (Lee et al., 2012d). For each individual disease a linear regression of ln(AAO) on expanded repeat length was performed, generating the coefficients given below.

Disease	Sample N	A	B	*p*
HD	445	6.120	-0.053	<2e-16
SCA1	177	5.683	-0.044	<2e-16
SCA2	294	5.799	-0.057	<2e-16
SCA3	397	7.137	-0.049	<2e-16
SCA6	69	5.967	-0.087	0.00268
SCA7	73	4.643	-0.026	2.94e-5
SCA17	7	2.387	0.017	0.7

Table 3.6. Effects of repeat length of the expanded allele on age at onset.
Results of fitting a linear regression ln(AAO) = A + B(CAG)n. p value refers to the significance of the regression parameter (B) indexing the effect of repeat length. HD = Huntington's disease; SCA = spinocerebellar ataxia.*

These parameters were used to calculate expected AAO for each individual based on their pathogenic repeat length. This was subtracted from their actual AAO to give a residual. The association of each SNP with AAO was tested by performing a linear regression of these residuals on the number of minor alleles in the genotype in PLINK (Purcell et al., 2007).

The primary analysis tested whether there was an overall association of all 22 SNPs with AAO. This was achieved by combining the association p values for each SNP using Brown (1975). Essentially, this is Fisher's method for combining p values corrected for linkage disequilibrium between SNPs.

The primary analysis used one-sided p values for association in the same direction as that observed in GeM-HD. In order to assess the overall directionality of the associations, the significance was compared to that obtained from a similar analysis using two-sided p values.

Analyses were performed on the following eight groups. Because of small sample size, SCA17 was not analysed independently, but was included in the analyses of all SCAs and HD+SCAs. p values were Bonferroni corrected for eight tests.

1. All polyglutamine diseases (HD+SCAs)
2. HD
3. All SCAs
4. SCA1
5. SCA2
6. SCA3
7. SCA6
8. SCA7

3.4 Contributions

The study was designed by Professors Tabrizi, Jones and Houlden. I was responsible for collection of phenotypic data on HD patients at the National Hospital for Neurology and Neurosurgery (London, UK). Data acquisition from other cohorts was performed by Conceicao Bettencourt, Davina Hensman, Sarah Wiethoff, Alexis Brice, Cyril Goizet, Giovanni Stevanin, Georgios Koutsis, Georgia Karadima, Marios Panas, Petra Yescas-Gomez, Lizbeth Esmerelda Garcia-Velazquez, Maria Elisa Alonso-Vilatela, Manuela Lima, Mafalda Raposo, Bryan Traynor and the SPATAX network. Statistical analyses were conducted by Professor Holmans. The work was supervised by Professors Holmans, Houlden, Tabrizi and Jones. These results were published in Bettencourt et al. (2016).

3.5 Results

In the primary analysis which grouped all 22 SNPs, and corrected for multiple comparisons, there was significant association with age at onset (AAO) in HD+SCAs (p = 1.43E-5), HD (p = 1.94E-3), all SCAs (p = 1.07E-3), SCA2 (p = 3.50E-3) and SCA6 (p = 1.62E-3). These associations were more significant than an undirected test using two-sided SNP p values, indicating the effect direction for all SNPs is consistent with the GeM-HD GWAS (GeM-HD, 2015). The association with HD AAO also replicates the GeM-HD results in an independent cohort.

Disease Group	GeM-HD concordance?	P (All SNPs)	P (High LD SNPs removed)	P (rs3512 removed)
ALL (HD+SCAs)	non directional	4.74E-04 *	2.26E-04 *	0.0049 *
	Same as GeM-HD	1.43E-05 *	6.98E-06 *	2.26E-04 *
HD	non directional	0.0226	0.0078	0.0364
	Same as GeM-HD	0.0019 *	0.0005 *	0.0039 *
SCAs	non directional	0.0188	0.0236	0.0771
	Same as GeM-HD	0.0011 *	0.0014 *	0.0067 *
SCA1	non directional	0.3760	0.3860	0.4440
	Same as GeM-HD	0.4160	0.2870	0.5240
SCA2	non directional	0.0230	0.0629	0.0233
	Same as GeM-HD	0.0035 *	0.0138	0.0044 *
SCA3	non directional	0.1760	0.1140	0.3550
	Same as GeM-HD	0.0809	0.0381	0.2050
SCA6	non directional	0.0059 *	0.0735	0.0051 *
	Same as GeM-HD	0.0016 *	0.0340	0.0016 *
SCA7	non directional	0.1550	0.2170	0.2970
	Same as GeM-HD	0.0447	0.0885	0.1130

Table 3.7. Analysis of combined SNPs.

*p values obtained by combining single-SNP p values using Brown (1975), allowing for LD between SNPs. Non-directional analysis combines two-sided p values. "Same as GeM-HD" analyses combine one-sided p-values in the same direction as the SNP effects observed in GeM-HD study(GeM-HD, 2015). In the "High LD SNPs removed" analysis, rs1037700, rs5893603 and rs16869352 were removed due to high LD (r2>0.8) with more significant SNPs in GeM-HD. * p values that satisfy Bonferroni correction for 8 disease group tests. Note that SCA17 was included in the "HD+SCAs" and "All SCAs" grouped analyses, but was not tested independently due to small sample size. HD – Huntington's disease; SCA – spinocerebellar ataxia*

In a secondary analysis, association of individual SNPs with AAO was examined. Three of these associations were significant in the same direction as GeM-HD after Bonferroni correction for eight disease combinations and 22 SNPs; rs3512 (*FAN1*) with all SCAs and HD+SCAs, and rs1805323 (*PMS2*) with HD+SCAs. Correcting for 22 SNPs, five associations were significant in the same direction as GeM-HD; rs1805323 (*PMS2*) in HD and SCA1 and in *RRM2B* rs1037699, rs1037700 and rs5893603 were associated with SCA6.

rs146353869, the most significant SNP in GeM was not replicated in this study, most likely because this small sample is less well powered to detect associations with rarer variants (rs146353869 MAF = 0.017). However, rs3512 is in strong linkage disequilibrium with the second GeM GWAS chromosome 15, associated with 1.4 later onset.

Three SNPs (rs1037700, rs5893603, and rs16869352) were in high LD ($r^2 > 0.8$) with more significant SNPs. Removing them reduced the combined SNP significance with SCAs, though they remained nominally significant (Table 6). Removing the most significant SNP in this study, rs3512, all combined SNP associations remained significant, suggesting the signal is not being driven by a single variant.

SNP	Chr	Pos	A1 (GeM-HD)	A2 (GeM-HD)	MAF (GeM-HD)	Beta (GeM-HD)	P (GeM-HD)	A1 (All)	A2 (All)	MAF (All)	Beta (All)	P (All)	Beta (HD)	p(HD)	Beta (SCA1)	P(SCA1)	Beta (SCA2)	P (SCA2)	Beta (SCA3)	P (SCA3)	Beta (SCA6)	P (SCA6)	Beta (SCA7)	P(SCA7)	Beta (AllSCA)	P (AllSCA)
rs1800937	2	48025764	T	C	0.092	0.820	4.30E-03	T	C	0.074	0.490	4.75E-01	0.520	6.21E-01	-0.571	6.51E-01	-0.459	8.18E-01	2.455	4.47E-02	0.614	8.25E-01	-10.050	5.34E-01	0.438	6.13E-01
rs4150407	2	128049631	C	T	0.444	0.575	4.60E-04	G	A	0.479	0.064	8.50E-01	-0.585	2.53E-01	-0.574	3.91E-01	1.384	1.03E-01	-0.013	9.85E-01	-2.129	2.55E-01	-2.702	3.83E-01	0.260	5.48E-01
rs5742933	2	190649316	C	G	0.206	-0.699	9.49E-04	C	G	0.205	-0.725	9.59E-02	-0.732	2.49E-01	1.102	2.19E-01	-2.333	3.69E-02	-1.005	2.19E-01	0.939	6.76E-01	0.551	8.77E-01	-0.714	2.03E-01
rs1799977	3	37053568	G	A	0.319	0.847	7.16E-07	G	A	0.280	-0.359	3.55E-01	0.531	3.39E-01	-0.241	7.58E-01	-2.555	2.20E-02	1.081	1.31E-01	-1.424	4.17E-01	-5.899	1.44E-01	-0.698	1.68E-01
rs6151792	5	80056961	T	C	0.099	-1.049	2.09E-04	T	C	0.117	-0.662	2.16E-01	-1.395	7.99E-02	-0.436	7.30E-01	0.495	6.79E-01	-1.347	2.21E-01	-1.577	5.07E-01	-0.116	9.77E-01	-0.350	6.09E-01
rs115109737	5	80102444	A	G	0.060	-1.289	4.50E-04	A	G	0.041	-2.095	1.81E-02	-3.014	2.10E-02	1.321	5.13E-01	-4.651	8.59E-02	0.100	9.50E-01	-5.197	1.41E-01	-5.756	2.87E-01	-1.726	1.28E-01
rs71636247	5	80118976	G	A	0.054	-1.398	2.55E-04	G	A	0.034	-2.208	2.63E-02	-1.917	1.89E-01	-1.974	4.39E-01	-6.400	4.86E-02	0.324	8.52E-01	-3.813	2.82E-01	-7.123	2.49E-01	-2.329	6.76E-02
rs1805323	7	6026942	T	G	0.038	-0.950	3.04E-02	A	C	0.043	-3.605	**3.14E-05****	-3.890	**3.14E-04***	-5.677	**1.67E-03***	-1.835	3.94E-01	-2.307	2.70E-01	-2.123	5.50E-01	-17.190	1.44E-01	-3.305	6.62E-03
rs12531179	7	6028687	T	C	0.147	0.938	3.84E-05	T	C	0.169	0.579	2.16E-01	1.070	1.23E-01	0.039	9.67E-01	1.137	3.08E-01	0.083	9.32E-01	-0.320	8.83E-01	-0.798	8.07E-01	0.367	5.39E-01
rs3735721	8	103217695	G	A	0.085	-1.529	5.68E-07	G	A	0.083	-0.389	5.25E-01	0.354	6.56E-01	0.692	5.13E-01	-3.278	6.32E-02	1.308	3.11E-01	-15.150	2.35E-03	-3.035	5.89E-01	-0.790	3.47E-01
rs1037700	8	103250775	C	G	0.097	-1.541	5.03E-08	G	C	0.094	-0.817	1.54E-01	-0.012	9.87E-01	1.046	2.46E-01	-4.132	2.11E-02	0.863	4.72E-01	-14.250	**5.47E-04***	-8.021	1.55E-01	-1.235	1.11E-01
rs5893603	8	103250839	G	-	0.097	-1.548	4.28E-08	G	-	0.093	-0.983	8.89E-02	-0.092	9.05E-01	0.914	3.13E-01	-4.189	1.84E-02	0.537	6.59E-01	-11.770	**2.13E-03***	-9.077	1.24E-01	-1.441	6.45E-02
rs1037699	8	103250930	T	C	0.096	-1.570	2.70E-08	A	G	0.094	-0.819	1.53E-01	-0.006	9.94E-01	0.758	4.13E-01	-3.519	3.97E-02	0.896	4.55E-01	-14.260	**4.86E-04***	-9.077	1.24E-01	-1.228	1.11E-01
rs16869352	8	103306033	C	T	0.083	-1.528	4.01E-07	C	T	0.080	-0.464	4.57E-01	0.691	3.98E-01	0.756	4.36E-01	-2.854	1.25E-01	0.681	6.27E-01	-10.850	3.24E-02	-7.745	1.64E-01	-1.067	2.09E-01
rs61752302	8	103311153	T	C	0.023	-1.671	3.03E-03	T	C	0.026	-0.150	8.92E-01	-0.520	7.46E-01	0.567	7.10E-01	-1.045	6.76E-01	4.882	1.18E-01	-8.015	1.69E-01	NA	NA	0.019	9.89E-01
rs72734283	14	75495059	G	A	0.099	0.858	4.32E-03	G	A	0.089	0.898	1.40E-01	2.057	1.14E-02	1.585	1.82E-01	-1.099	5.41E-01	-0.650	5.88E-01	-1.686	5.59E-01	10.770	3.82E-02	0.318	6.98E-01
rs175080	14	75513828	A	G	0.466	-0.434	7.72E-03	A	G	0.435	-0.671	5.66E-02	-1.245	1.61E-02	0.279	7.16E-01	-0.090	9.23E-01	0.397	5.62E-01	-0.927	5.84E-01	-4.356	1.66E-01	-0.405	3.70E-01
rs146353869	15	31126401	A	C	0.017	-6.107	4.30E-20	A	C	0.017	-2.362	8.17E-02	-1.804	3.28E-01	1.980	5.64E-01	-8.999	3.81E-02	-1.537	4.94E-01	-3.496	5.52E-01	7.338	6.60E-01	-2.610	1.48E-01
rs114136100	15	31197976	T	C	0.018	-5.073	8.49E-16	T	C	0.019	-2.101	9.20E-02	-1.188	4.88E-01	1.609	6.00E-01	-1.168	7.89E-01	-3.519	8.25E-02	-3.464	5.55E-01	6.909	6.73E-01	-2.521	1.27E-01
rs150393409	15	31202961	A	G	0.016	-5.765	9.34E-18	A	G	0.013	-2.735	7.03E-02	-2.909	1.39E-01	-0.354	9.28E-01	-4.224	4.88E-01	-3.176	1.92E-01	-0.912	8.99E-01	7.443	6.57E-01	-2.551	2.17E-01
rs3512	15	31235005	C	G	0.309	1.325	5.28E-13	G	C	0.283	1.680	**1.52E-05****	1.297	2.94E-02	1.388	8.70E-02	1.020	3.03E-01	2.156	2.36E-03	0.886	6.37E-01	9.647	5.00E-03	1.809	**2.22E-04****
rs20579	19	48668830	A	G	0.124	0.769	6.65E-03	T	C	0.134	0.427	4.09E-01	0.119	8.82E-01	1.244	2.84E-01	0.412	7.55E-01	1.099	2.17E-01	-7.791	2.19E-02	-0.216	9.54E-01	0.515	4.28E-01

Table 3.8. Individual SNP association with age at onset.

Beta denotes the effect size – that is, the number of years added to or subtracted from the expected age at onset for each copy of the minor allele (A1). MAF denotes the frequency of the minor allele in GeM-HD(GeM-HD, 2015) (Column 6) and the present study (Column 11). p values highlighted bold and "" satisfy Bonferroni correction for 22 SNPs; those highlighted bold and "**" satisfy Bonferroni correction for 8 disease groups and 22 SNPs. Note that for SNPs in reverse orientation (rs4150407, rs1805323, rs1037700, rs1037699, rs3512, and rs20579) are complementary to those in GeM-HD, which uses HGVS nomenclature.*

A polygenic score was calculated using the sum of minor alleles at each locus, weighted by their effect size in the GeM GWAS. There was a positive correlation between increasing polygenic score and residual AAO, suggesting the variants are associated with delayed onset. However, the effect was relatively small, accounting for approximately 1% of the variance in residual AAO.

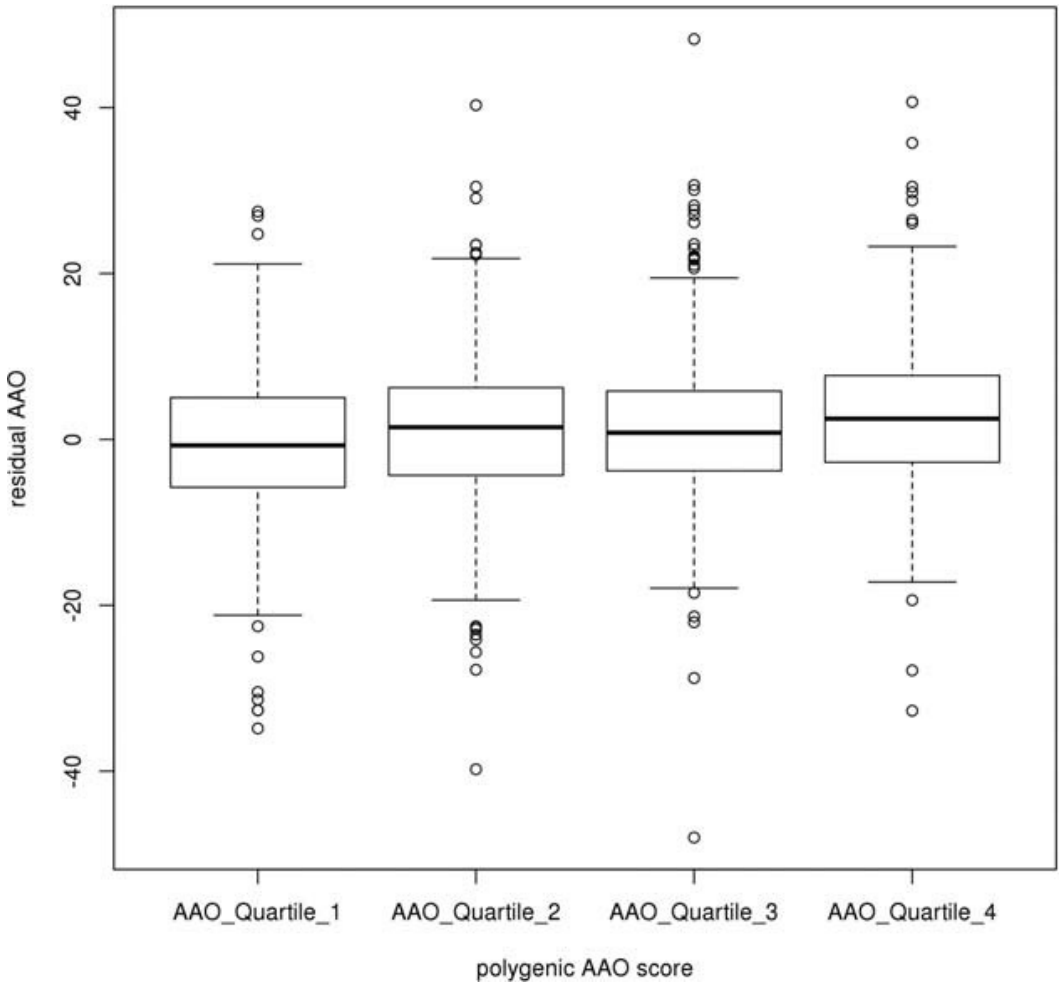


Figure 3.1. Boxplot of residual age at onset across all samples against polygenic score.
Polygenic score was calculated by summing minor alleles (weighted by effect on AAO in the GeM GWAS). Residual AAO for each quartile of risk score is plotted. Lower scores correspond to earlier than expected AAO (smaller residuals).

3.6 Discussion

These data suggest a common mechanism operates in polyglutamine diseases in which variation in DNA repair genes modifies age at onset (AAO). The most significant variant in the present study, rs3512, is a common variant (MAF = 0.19) in the 3' untranslated region (UTR) of *FAN1* and has not previously been linked with disease. It was associated with 1.31 year delayed onset in GeM (p = 5.28E-13). FAN1 is a DNA endo- and exonuclease involved in DNA repair (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Smogorzewska et al., 2010a) that is highly expressed in the brain (GTEx, 2015). It was originally identified as a key component in the Fanconi anaemia (FA) interstrand cross-link (ICL) repair pathway, though its mutation does not cause Fanconi anaemia.

Significant association was also seen with *PMS2* and *RRM2B*. PMS2 is a nuclease that heterodimerises with MLH1 to form MutLα. After MMR initiation by MutSα (MSH2-MSH6) or MutSβ (MSH2-MSH3) binding to a mismatch, MutLα introduces a single strand break (SSB) near the mismatch, thereby generating new entry point for the exonuclease EXO1 to degrade the mismatch (Kadyrov et al., 2006, Borras et al., 2013). MutLα may also recruit DNA polymerase III to the mismatch site. *PMS2* mutations are known to cause dominant and recessive hereditary bowel cancer (Online Mendelian Inheritance in Man, 2015). Its knockout in myotonic dystrophy and Friedreich's ataxia mouse models significantly reduced repeat expansion (Gomes-Pereira et al., 2004, Bourn et al., 2012). *RRM2B* supplies dNTPs for DNA repair. It is widely expressed, including throughout the nervous system (GTEx, 2015), and its mutation can cause mitochondrial DNA depletion syndromes such as autosomal dominant progressive external ophthalmoplegia (McKusick, 2007), but it has not been linked with repeat instability.

DNA repair has been implicated in several rare genetic diseases, all of which share a similar phenotype of neurodegeneration in the cerebellum or basal ganglia (Ross and Truant, 2017). Rare loss of function mutations in DNA repair genes are known to cause several recessive ataxias. ATM is a master regulator of DNA repair following double strand breaks. PNPK is a DNA-specific kinase involved in DNA repair. APTX interacts with PARP1 in the repair of single strand breaks. TDP1 mutations produce defects in single-strand break repair. It is not clear how these deficits in DNA repair result in cerebellar degeneration. However, there is evidence for ATM control being critical in cell division and apoptosis, which could lead to neuronal loss (Shiloh and Ziv, 2013).

Repetitive DNA sequences form abnormal secondary structures, such as hairpin loops, which may be substrates of DNA repair proteins, and the process of repair may result in somatic instability (Holmans et al., 2017). MMR proteins, including MSH3, bind directly to slipped-strand DNA structures formed by CAG repeats (Pearson et al., 1997), processing them in complex with MSH2 and PMS2. Larger CAG repeats are associated with more severe pathology, earlier onset and faster progression, so somatic expansion provides a plausible mechanism by which DNA repair variation may act (Massey and Jones, 2018).

Alternatively, aberrant DNA repair may lead to the accumulation of oxidative DNA damage in neurons, leading to mutation and epigenetic modification that alters expression. The nervous system is particularly vulnerable to oxidative stress because it has a relatively high metabolic rate, generates high levels of reactive oxygen species and is thought to have a lower ratio of anti- to pro-oxidant enzymes (Canugovi et al., 2013). Over time, these changes could lead to

neuronal dysfunction and ultimately, once a threshold level is crossed, apoptosis and neuronal loss could be triggered (Ross and Truant, 2017).

This study had several limitations, including the relatively small sample size for several rare ataxias such as SCAs 6, 7 and 17. Though the relationship between CAG length and AAO was modelled independently for each disease, there are likely to be other factors which differ between diseases that could not be accounted for, such as the presence of interruptions in the repeat tract which could provide protective stabilisation (Menon et al., 2013). Interruptions have been demonstrated in HD, SCAs 1-3 and 17, fragile X syndrome and Friedreich's ataxia and myotonic dystrophy (Massey and Jones, 2018). They are known to reduce the stability of abnormal hairpin structures, thereby reducing repeat expansion. The next step is to repeat the analysis with larger samples and more DNA repair variants as this would be expected to increase the predictive power of the polygenic risk score (Dudbridge, 2013, Purcell et al., 2009, Consortium, 2014).

3.7 Summary

DNA repair genes modify age at onset (AAO) in polyglutamine diseases as a group, as well as independently in all SCAs, HD, SCA1 and SCA6, though the effect size is relatively small. This study replicates the results of the GeM GWAS in an independent cohort, with significant association of variation in FAN1 on chromosome 15, PMS2 on chromosome 7, and RRM2B on chromosome 8 with AAO. These results suggest a common mechanism operates in polyglutamine diseases, meaning therapeutic opportunities may be relevant across multiple diseases. Molecules targeting DNA repair have been developed to treat cancer. Such therapies may prove beneficial in polyglutamine diseases (Farmer et al., 2005, Jackson and Chester, 2015). A polygenic risk score may improve clinical trial design by enabling stratification of patients by genetic variability. When the pathogenic trinucleotide repeat expansions causing these diseases were first discovered, the research focus was on the mutation itself; having subsequently explored an array of downstream pathogenic mechanisms, focus is now returning to the DNA level.

3.8 Publications relating to this chapter

The work relating to this chapter was published in:

DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Bettencourt, C.*, Moss, D. H.*, **Flower, M.***, Wiethoff, S., Brice, A., Goizet, C., Stevanin, G., Koutsis, G., Karadima, G., Panas, M., Yescas-Gomez, P., Garcia-Velazquez, L. E., Alonso-Vilatela, M. E., Lima, M., Raposo, M., Traynor, B., Sweeney, M., Wood, N., Giunti, P., network, Spatax, Durr, A., Holmans, P.[#], Houlden, H. [#], Tabrizi, S. J. [#] and Jones, L. [#] ***Ann Neurol***, 2016 Jun;79(6):983-90. doi: 10.1002/ana.24656.

* These authors should be regarded as joint first authors.

[#] These authors jointly supervised the work.

Chapter 4 Transcriptional dysregulation in Huntington's disease patient blood

4.1 Background

4.1.1 Huntington's disease causes widespread pathology

HD research has traditionally focused on the brain due to the presence of characteristic mutant huntingtin protein aggregates (Bates et al., 2015c) and because the prominent symptoms and signs can be linked to neurodegeneration in the basal ganglia and cerebral cortex (van der Burg et al., 2009). However, mutant *HTT* is ubiquitously expressed (Trottier et al., 1995) and mounting evidence suggests it has direct effects in peripheral tissues (van der Burg et al., 2009, Carroll et al., 2015), though whether these effects are distinct, or parallel those in the brain remains unclear.

HD patients demonstrate peripheral immune dysfunction presymptomatically (Tai et al., 2007a, Bjorkqvist et al., 2008, Kwan et al., 2012c, Träger et al., 2015), as well as weight loss that leads to cachexia with advancing disease (Carroll et al., 2015). There is progressive muscle wasting (Busse et al., 2008), endocrine dysfunction (Saleh et al., 2009), liver impairment (Carroll et al., 2015) and cardiac dysfunction (Lanska et al., 1988, Mihm et al., 2007, Pattison et al., 2008). Mutant HTT (mHTT) protein aggregates can be found in the peripheral tissues of HD mice (Orth et al., 2003), as well as advanced patients (Turner et al., 2007). These peripheral features may contribute to CNS pathology, disease progression and mortality (Carroll et al., 2015, van der Burg et al., 2009), and strongly suggest that HD is a systemic disorder.

Patient-derived tissue is the most physiologically relevant system in which to study pathogenesis. The peripheral phenotype provides an opportunity to study HD pathogenic mechanisms. In contrast to brain tissue, availability of which is limited and from post-mortem subjects with end-stage disease (Montanini et al., 2013, Tomita et al., 2004), peripheral tissues can be sampled minimally invasively and inexpensively from living patients, enabling longitudinal study throughout disease course.

4.1.2 Transcriptional dysregulation in Huntington's disease

4.1.2.1 Transcriptional dysregulation in the brain

Transcriptional dysregulation is an early and central feature of HD (Seredenina and Luthi-Carter, 2012, Hodges, 2006). Suggested mechanisms include the sequestration of transcription factors in intracellular aggregates or their direct inhibition by mutant HTT (mHTT). The transcription factor CBP, for example, is found in polyglutamine aggregates (McCampbell et al., 2000, Nucifora et al., 2001), whereas Sp1 has a higher affinity for mutant than wild type HTT (Dunah et al., 2002, Li et al., 2002, Shimohata et al., 2000). Expansion of the polyglutamine stretch in HTT reduces its affinity for the transcriptional repressor REST, which the wild type protein retains in the cytoplasm, resulting in repression of neuronal genes such BDNF (Zuccato et al., 2007, Zuccato et al., 2003). mHTT can also affect proteasomal degradation of transcription factors, decreasing levels of CBP, for example (Cong et al., 2005). Transcriptional activity is influenced by chromatin structure; cell and animal models have found histone hypoacetylation in HD (Sadri-Vakili et al., 2007, Hazeki et al., 2002), and HDAC inhibitors have been shown to be neuroprotective, improving the HD-like phenotype (Ferrante et al., 2003, Sadri-Vakili et al., 2007, Steffan et al., 2001, Hockly et al., 2003, Thomas et al., 2008). HTT itself can bind DNA, altering the activity of transcription factors (Benn et al., 2008b), and is found in P-bodies, which are involved in RNA-

mediated transcriptional regulation (Eulalio et al., 2007, Savas et al., 2008). miRNA expression is dysregulated in HD cortex (Johnson et al., 2008, Packer et al., 2008, Johnson and Buckley, 2009, Lee et al., 2011b, Strand et al., 2005), with notable reduction of those regulated by REST (Marti et al., 2010, Packer et al., 2008). By grouping genes into biologically relevant pathways several process have been reproducibly implicated, including neuronal signalling, synaptic proteins, transcription factors, chromatin remodelling, metabolic regulators and inflammation (Seredenina and Luthi-Carter, 2012).

Neueder and Bates (2014) applied weighted gene correlation network analysis (WGCNA) (Langfelder and Horvath, 2008) to the Hodges et al. (2006) microarray brain expression data set of 44 human HD and 36 matched control brains. They generated networks for four brain regions; the caudate nucleus (CN), BA4 region of the frontal cortex, which has motor function (FC-BA4), BA9 region of the frontal cortex, involved in association and cognitive functions (FC-BA9), and cerebellum (CB).

In the cerebellum they identified 2504 downregulated and 2230 upregulated genes. The modules CBpos4, CBpos5 and CBneg4 were also dysregulated in frontal cortex, and CBpos5 and CBneg2 were dysregulated in both frontal cortex and caudate. CBneg1 hub genes were involved in synaptic function. Other negatively correlated cerebellar modules enriched for mitochondrial and proteasomal genes. Positively correlated cerebellum modules enriched for genes involved in transcriptional regulation, chromatin binding and protein folding. Notably, there was no immune dysregulation in the cerebellum.

In the frontal cortex, there was no transcriptional dysregulation in Brodmann area 9 (dorsolateral and medial prefrontal cortex), but significant changes were found in Brodmann area 4 (primary motor cortex). Most modules dysregulated in BA4 were also dysregulated in caudate and cerebellum. In addition to the signal shared with cerebellum, there was upregulation of inflammatory response and NFκB/IκB genes, and angiogenesis, and there was downregulation of synaptic, mitochondrial, protein transport and proteasomal genes.

The largest changes were seen in the caudate, with 3798 (30.4%) of genes negatively and 5349 (42.8%) of genes positively dysregulated. Caudate modules were better preserved in the frontal cortex than the cerebellum. Positively correlated modules were involved in the inflammatory response, transcriptional regulation and mRNA processing, and negatively correlated ones in synaptic, mitochondrial, DNA repair and proteasomal function. They found preservation of caudate modules in other neurodegenerative diseases, including AD, ALS, MS, PD and myotonic dystrophy, with a high enrichment for inflammatory pathway genes. HD mouse models mimicked some of the transcriptional dysregulation, but aspects such as the inflammatory response were poorly reflected.

Labadorf et al. (2015) analysed the transcriptome of human postmortem prefrontal cortex Brodmann area 9 (BA9) from 20 HD subjects and 49 controls using next-generation high throughput sequencing, identifying dysregulation of immune and developmental genes. They found significant differential expression in HD including proinflammatory genes of the NFkB family and cytokine receptors. Dysregulated pathways enriched for genes involved in the immune response, development, cell growth and transcriptional regulation. However, none of the expression changes were significantly associated with disease burden or age at onset.

4.1.2.2 Peripheral transcriptional dysregulation

In R6/2 mice, brain and muscle have been shown to be concordant (Luthi-Carter et al., 2002, Strand et al., 2005), and this profile may be consistent with HD patient muscle too (Strand et al., 2005). However, studies of gene expression changes in HD blood have been inconsistent. Using microarray technology, Borovecki et al. (2005) identified 12 upregulated transcripts, seven of which were also upregulated in brain. However, subsequent studies did not replicate these results (Runne et al., 2007, Lovrecic et al., 2009, Mastrokolias et al., 2015).

Gene names	Protein	Function
ANXA1	Annexin A1	Inflammatory regulator
AXOT	Axotrophin	Ubiquitin-protein ligase
CAPZA1	Capping protein (Actin filament) muscle Z-line, alpha 1	Structural protein that binds actin filaments
HIF1A	Hypoxia-inducible factor 1, alpha subunit	Transcriptional regulator of the adaptive response to hypoxia
JJAZ1 (SUZ12)	Polycomb protein SUZ12	Transcriptional repressor
P2Y5 (LPAR6)	Lysophosphatidic Acid Receptor 6	G-protein coupled receptor
PCNP	PEST proteolytic signal-containing nuclear protein	Cell cycle regulation
ROCK1	Rho-associated protein kinase 1	Protein kinase that regulates actin cytoskeleton
SF3B1	Splicing factor 3b, subunit 1	Pre-mRNA splicing
SP3	Transcription factor Sp3	Transcription factor
TAF7	Transcription initiation factor TFIID subunit 7	Transcription factor
YIPPEE (YPEL5)	Yippee Like 5	Innate immune system

Table 4.1. 12 genes significantly upregulated in HD blood from Borovecki et al. (2005).

Using tag-based serial analysis of gene expression (SAGE), Mastrokolias et al. (2015) found 170 genes differentially expressed by motor score, 40 of which had previously been reported in at least one microarray study.

Gene	Protein	B	Expression level	Adjusted p	Protein Function
HYAL2	Hyaluronoglucosaminidase 2	0.4	2.6	1.0E−03	Hydrolyzes hyaluronic acid
LMO2	LIM domain only 2	0.3	6.6	1.0E−03	Yolk sac hematopoiesis
MARC1	Mitochondrial amidoxime reducing C1	0.4	5	5.0E−03	N-hydroxylate prodrug conversion
NT5DC2	5′-Nucleotidase domain containing 2	0.4	2.8	9.0E−03	Hydrolase and metal ion binding
RNF135	Ring finger protein 135	0.3	5.8	9.0E−03	DDX58 Ubiquitination~IFN-β
PROK2	Prokineticin 2	0.5	7.9	1.0E−02	Circadian clock—GI contraction
RPN1	Ribophorin I	0.3	5.5	1.0E−02	26S Proteasome ubiquitin binding
CYSTM1	Cysteine-rich transmembrane module 1	0.4	6	1.0E−02	Stress tolerance
VCAN	Versican	0.3	8.2	1.6E−02	Intercellular signaling Binds hyal. acid
NCF4	Neutrophil cytosolic factor 4	0.3	8.9	1.8E−02	NADPH-oxidase component
ARL4C	ADP-Ribosylation factor-like 4C	−0.3	8.2	1.0E−03	Microtubule vesicular transport
TMEM109	Transmembrane protein 109 (Mg23)	−0.3	7	6.0E−03	UVC αB-Crystallin protection
MACF1	Microtubule-actin crosslinking factor 1	−0.2	7.2	6.0E−03	Actin-microtubule stabilization
MDN1	Midasin homolog	−0.2	5.3	7.0E−03	AAA-ATPase(dynein)
PTPN4	Protein tyrosine phosphatase NR type 4	−0.3	5.1	9.0E−03	Glutamate receptor signaling
PRF1	Perforin 1	−0.4	9.5	1.0E−02	Cytolysis
CD3G	CD3g Molecule gamma	−0.3	7.5	1.0E−02	CD3 complex signal transduction
NMT2	N-Myristoyltransferase 2	−0.3	3.4	1.0E−02	N-terminal Myristoylation
KLRD1	Killer cell lectin receptor subfamily D 1	−0.4	6.1	1.0E−02	Recognition of MHC class I HLA-E
GPR56	G Protein-coupled receptor 56	−0.4	7.3	1.0E−02	Brain cortical patterning

Table 4.2.Top 10 up and downregulated genes in HD blood from Mastrokolias et al. (2015).

B – coefficient of gene expression change per motor score unit, multiplied by average motor score. Expression level – average log2 gene expression.

Miller et al. (2016) identified transcriptional dysregulation in HD monocytes. Patient-derived primary monocytes were cultured with and without proinflammatory stimulus. In their basal, unstimulated state there was significant transcriptional dysregulation, including increased expression of proinflammatory cytokines such as IL-6. Their pathway analysis enriched for proinflammatory genes regulated by the NFκB pathway. These results suggest peripheral HD

myeloid cells have a proinflammatory phenotype in resting state, consistent with them being primed by mHTT. This leads to an exaggerated inflammatory response once they encounter a stimulus.

4.1.3 RNA-Seq

Prior to next-generation sequencing (NGS), large scale gene expression was studied with microarrays. Thousands of DNA probes profiled transcripts, but gave limited coverage of only known and common alleles, and lacked coverage of variant transcripts. RNA-Seq utilises NGS and is not dependent on prior sequence knowledge, sequencing every transcript in the sample, known and unknown. This allows the identification of structural variations, such as gene fusions and alternative splicing, as well as novel genes and transcripts. It is more sensitive at detecting low abundance transcripts, can differentiate isoforms, and accurately determines expression level, differential splicing and allele-specific expression. It gives an absolute measure of transcript levels and structure (Wang et al., 2009, Blekhman et al., 2010).

4.2 Aim

Analyse the whole blood transcriptome of a cohort of Huntington's disease subjects using RNA sequencing (RNA-Seq), correlate this with disease severity and compare it to transcriptional dysregulation seen in the brain in HD and other neurodegenerative diseases.

4.3 Methods

4.3.1 Cohorts

The Track-HD cohort consisted of 54 premanifest gene carriers, 63 manifest HD subjects and 23 controls. These were selected as a representative sample from the Track-HD study to ensure a wide range of disease risk and severity. Control subjects were age and gender matched to individuals in the premanifest and manifest groups, and selected from spouses or partners to ensure consistency of environments. Track-HD enrolled participants at four study sites in London (UK), Paris (France), Leiden (Netherlands), and Vancouver (BC, Canada) (Tabrizi et al., 2009b). *Manifest* subjects demonstrated motor abnormalities that were unequivocal signs of HD, as evidenced by total motor scores (TMS) over 5 on the Unified Huntington's Disease Rating Scale (UHDRS). *Premanifest* gene carriers had a burden of pathology score (age x [CAG – 36.5]) (Penney et al., 1997) greater than 250, a TMS of 5 or lower and a diagnostic confidence score (DCS) less than 4 on the UHDRS (Group, 1996), indicating no substantial motor signs (Tabrizi et al., 2009b). Age and clinical scores considered for the analysis were at the time of blood collection.

The Leiden cohort (Mastrokolias et al., 2015) consisted of 18 premanifest gene carriers, 56 manifest HD subjects and 27 age and gender-matched controls. Motor onset was determined by an experienced neurologist using the same UHDRS standard as in TRACK-HD. All premanifest carriers showed no substantial motor signs, with a TMS of 5 or less and a UHDRS diagnostic confidence level less than 4. All controls were free of known medical conditions. Blood sample collection and analysis methods, described below, were identical for the two cohorts.

Cohort	Group	n	Mean age, y ± SD (range)	Gender (male/female)	Mean (CAG)n length ± SD (range)	Mean TMS ± SD (range)	Mean TFC ± SD (range)
Track-HD	*Premanifest*	*50*	*42 ± 9 (22-64)*	*24/26*	*43 ± 3 (39-52)*	*2 ± 2 (0-8)*	*13 ± 0 (12-13)*
	Manifest	*62*	*48 ± 10 (23-64)*	*26/36*	*44 ± 3 (39-59)*	*23 ± 11 (5-45)*	*11 ± 2 (7-13)*
	HD	112	46 ± 10 (22-64)	50/62	44 ± 3 (39-59)	14 ± 13 (0-45)	12 ± 2 (7-13)
	Control	22	45 ± 5 (34-53)	9/13	-	-	-
Leiden	*Premanifest*	*18*	*46 ± 10 (29-63)*	*5/13*	*42 ± 2 (39-47)*	*3 ± 2 (0-5)*	*12 ± 1 (10-13)*
	Manifest	*56*	*55 ± 11 (35-79)*	*29/27*	*44 ± 3 (39-53)*	*42 ± 30 (6-102)*	*7 ± 5 (0-13)*
	HD	74	53 ± 11 (29-79)	34/40	44 ± 3 (39-53)	32 ± 31 (0-102)	8 ± 5 (0-13)
	Control	27	43 ± 11 (26-65)	13/14	-	-	-
Combined	HD	186	48 ± 11 (22-79)	84/102	44 ± 3 (39-59)	21 ± 24 (0-102)	10 ± 4 (0-13)
	Control	49	44 ± 9 (26-65)	22/27	-	-	-

Table 4.3. Track-HD and Leiden cohorts for RNA-Seq analysis.

Manifest subjects demonstrated motor abnormalities that were unequivocal signs of HD. Premanifest gene carriers had a total motor score of 5 or lower and a diagnostic confidence score (DCS) less than 4 on the UHDRS, indicating no substantial motor signs. The HD group consists of the combined premanifest and manifest subjects. Controls were matched for age and gender. Age and clinical scores considered for the analysis were at time of blood collection. SD – standard deviation; TFC – Total Functional Capacity; TMS – Total Motor Score.

4.3.2 Sample collection

Whole blood was collected in two PAXGene Blood RNA tubes (PreAnalytix, Qiagen/BD Company) per subject, and immediately placed upright at room temperature. They were checked at 5 hours for incomplete mixing or separation, and any showing separation were remixed with a further 10 inversions. Tubes were stored overnight at -20°C and transferred to -80°C the following morning. They were sent on dry ice to Biorep within 30 days.

4.3.3 RNA preparation

Total RNA extraction was performed using the PAXGene Blood RNA kit (cat #762174; PreAnalytix, Qiagen/BD Company), following the supplier's instructions. Each solution in the kit was divided into aliquots to process batches of 12 samples. Replicate tubes for each subject were processed on different days. RNA was stored at -80°C before proceeding with the quality measurements and further use. RNA was collected by centrifugation, washing with 70% ethanol, and resuspended in buffer. Quality measurements of total RNA were made using spectrophotometric analysis (Nanodrop), 260/280 ratio denaturing agarose gel, and the RNA 6000 Nano kit for the Agilent Bioanalyzer (cat # 5067-1511, Agilent Technologies). Samples were globin reduced using the GLOBINclear™ method (cat #AM1980, ThermoFisher Scientific). Quality control measures were made on globin-reduced samples on the Bioanalyzer RNA 6000 Nano kit (cat #5067-1511, Agilent Technologies).

4.3.4 Sequencing

Indexed cDNA sequencing libraries were prepared using the TruSeq™ Poly-A mRNA method (Illumina). In short, poly-A mRNA transcripts were captured from total RNA using poly-T beads and cDNA generated using random hexamer priming (Illumina, 2014). Paired-end sequencing of indexed cDNA libraries on a HiSeq 2500 generated at least 50 M reads per sample. Sequencing was performed using SBS and cluster kits from Illumina. Indexed samples were demultiplexed and FASTQ files were generated.

4.3.5 Quality control

Sequencing failed for six Track-HD samples, including four premanifest, one manifest and one control subject. Quality control analysis was performed using the RNA-SeQC package (DeLuca et al., 2012), ensuring measures including rRNA rate, mapping rate, concordance mapping rate and uniqueness rate were within acceptable ranges. Globin depletion was checked by inspecting read counts mapped to HBB, HBA1 and HBA2, confirming they made up less than 2% of reads for

all samples. Four Track-HD and six Leiden samples failed quality control for duplication rate over 75%, GC bias or 5' bias, and were removed, leaving 48 premanifest, 61 manifest and 21 control subjects in the Track-HD cohort and 15 premanifest, 54 manifest and 26 control subjects in the Leiden cohort.

4.3.6 Gene expression analysis

RNA-Seq data were aligned to the human reference genome hg19 using TopHat2 (Kim et al., 2013). Read counts were summarized using HTSeq, keeping any duplicates and using the Ensembl transcript/gene database (http://www.ensembl.org/info/data/ftp/index.html, obtained in gtf format, genome build GRCh38.3, gene build updated in June 2015). To remove residual batch effects the R package svaseq was used (Leek, 2014). Using the cleaned count data, differential expression analysis was conducted using the R package DESeq2 (Love et al., 2014). Outlier counts were removed using a Cooks distance cutoff of 5 in DESeq2. After filtering by the mean of normalised counts, 18,257 transcripts were detected. Age and gender were used as covariates in the analysis.

4.3.7 Pathway analysis

Enrichment of differential expression among gene sets corresponding to biological hypotheses (pathways) was tested using the Gene Set Enrichment Analysis (GSEA) method (Subramanian et al., 2005). Rather than defining a list of significant genes, GSEA ranks all genes in order of their differential expression statistic, and tests whether the genes in a particular gene set have a higher rank overall than would be expected by chance. The analysis is weighted by the differential expression statistic, thus giving more weight to more significant genes. Significance of enrichment was obtained by randomly permuting gene-wide association statistics among genes. One-sided p-values were calculated separately for differential upregulation and downregulation of expression in HD, and these were then converted into the corresponding chi-square statistic for use in the GSEA analysis. To avoid making a priori assumptions, a large pathway set was collated from publicly available pathway databases, including Gene Ontology (GO) (Consortium, 2016), Kyoto Encyclopedia of Genes and Genomes (KEGG) (KEGG, 2016), Mouse Genome Informatics (MGI) (MGI, 2016), PANTHER (PANTHER, 2016), BioCarta (BioCarta, 2016), REACTOME (REACTOME, 2016) and NCI (Institute, 2012). This resulted in a total of 14,706 functional gene sets, many with overlapping members, containing between 3 and 500 genes. To correct for multiple testing of pathways, GSEA p-values were converted into q-values (Storey and Tibshirani, 2003), which can be interpreted as the minimum false discovery rate at which that q-value would be counted as significant.

4.3.8 Gene co-expression networks

Weaknesses of relying on public databases to provide pathways for analysis include their restriction to prior biological knowledge and the poor annotation of many genes. To overcome this annotation gap, we also tested the following sets of gene co-expression modules for enrichment of dysregulation:

1. The set of 124 HD brain expression modules derived by Neueder and Bates (2014), who applied weighted gene correlation network analysis (WGCNA) (Langfelder and Horvath, 2008) to the Hodges et al. (2006) microarray brain expression data set of 44 human HD and 36 matched control brains. They generated networks for four brain regions; the caudate nucleus (CN), the BA4 region of the frontal cortex, which has motor function (FC-BA4), the BA9 region of the frontal cortex, involved in association and cognitive functions (FC-BA9), and cerebellum (CB).

2. A set of 117 co-expression modules derived from the Gibbs et al. (2010) dataset, comprising microarray expression data from 150 control individuals measured in four brain regions: cerebellum (CB), frontal cortex (FC), caudal pons (Pons) and temporal cortex (TCTX). Modules were generated using WGCNA as described in (International Genomics of Alzheimer's Disease, 2015).
3. A novel set of 213 co-expression modules were generated from Braineac (2016), which consists of microarray expression data for 12 brain regions from 134 control brains; occipital cortex, frontal cortex, temporal cortex, hippocampus, intralobular white matter, cerebellar cortex, thalamus, putamen, substantia nigra, and medulla (inferior olivary nucleus). For each brain region, the array data was normalised in the R statistical-programming environment using the RMA algorithm (Carvalho and Irizarry, 2010). Principal Component Analysis (PCA) and hierarchical clustering were used to identify single outlier arrays for removal. In addition, small outlier clusters (<6 arrays) that were distinct from most of the other arrays were removed (i.e. small clusters appearing at the top of the dendrogram). Once outlier arrays were removed, the arrays were re-normalised and inspected again and re-processed if necessary until a homogenous dataset was produced. WGCNA was performed using the R package to derive modules (Langfelder and Horvath, 2008). Multiple probesets of the same gene were collapsed to a single value using the collapseRows() function, using default settings and based on gene annotation provided by Affymetrix (Affymetrix, 2016). Scale independence and mean connectivity were plotted to derive a soft threshold power of 6. Networks were unsigned.
4. The set of 111 co-expression modules from Zhang et al. (2013), generated using microarray expression data on 1,647 postmortem samples from three brain regions of late-onset Alzheimer's disease (LOAD) and control subjects; prefrontal cortex (BA9), primary visual cortex (BA17), and cerebellum.

4.3.9 Concordance of fold change in gene expression between datasets

Labadorf et al. (2015) analysed the transcriptome of human postmortem prefrontal cortex Brodmann area 9 (BA9) from 20 HD subjects and 49 controls using next-generation high throughput sequencing, identifying dysregulation of immune and developmental genes. Of the 15,834 genes common to both the combined Track-HD and Leiden blood dataset and the Labadorf et al. (2015) prefrontal cortex dataset, 8447 had a fold change >1 (i.e. upregulated) in blood and 7860 in cortex. Thus, if fold changes in the two datasets were assumed to be unrelated, the expected probability that a gene would show concordant fold change is equal to ((8447/15834)x(7860/15834))+((7387/15834)x(7974/15834)) = 0.4997. The number of genes with concordant fold change in the absence of a relationship between the datasets is thus distributed as a binomial (15834, 0.4997) distribution. In the actual data, 8425 genes were observed to have concordant direction of fold change, significantly higher than the number expected by chance (7912).

A similar method was used to test for concordance of fold change in genes between the Track-HD and Mastrokolias et al. (2015) datasets.

4.4 Contributions

The study was conceived and designed by Professor Sarah Tabrizi. RNA was prepared by Davina Hensman and sequenced at Biorep. Alignment and read counts were performed by Kitty Lo and Vincent Plagnol. Gene expression and pathway analyses were performed by Professor Peter Holmans. Data interpretation and presentation was conducted by Michael

Flower. Figures were prepared by Timothy Stone and Michael Flower. Michael Flower wrote the manuscript, Hensman Moss et al. (2017a).

4.5 Results

4.5.1 Differential expression

Attempting to identify both HD specific and stage-specific changes in gene expression (mRNA) level, premanifest, manifest and control subjects were compared, whilst controlling for age and gender. *Premanifest* gene carriers had a mean total motor score (TMS) of 2 and total functional capacity (TFC) of 13, indicating no substantial motor signs. *Manifest* subjects demonstrated motor abnormalities that were unequivocal signs of HD.

No transcripts were significantly differentially expressed (FDR < 0.05) between premanifest and manifest HD in either the TRACK-HD (Tabrizi et al., 2009b) or the independent Leiden cohort, or when these cohorts were combined (results not shown). As expression changes did not differ significantly between disease stages, all mutant *HTT* gene carriers were combined to increase the analytical power in a comparison of HD and controls. Once again there were no individually significant transcripts in independent or combined cohorts. The top 10 genes from the differential expression analysis in the combined cohort are given below, and a complete list is provided in Hensman Moss et al. (2017a).

Entrez gene ID	Gene Symbol	p (diffexp)	q (diffexp)	log2(FC)
722	C4BPA	7.81E-06	1.41E-01	1.371
2297	FOXD1	9.09E-05	7.02E-01	-0.785
3805	KIR2DL4	1.93E-04	7.02E-01	0.651
196394	AMN1	2.11E-04	7.02E-01	0.208
94137	RP1L1	2.47E-04	7.02E-01	-1.350
158248	TTC16	2.67E-04	7.02E-01	-0.347
100422824	MIR3128	2.86E-04	7.02E-01	0.930
5797	PTPRM	3.12E-04	7.02E-01	-0.359
84692	CCDC54	4.79E-04	9.58E-01	2.532
889	KRIT1	7.30E-04	9.58E-01	-0.081

Table 4.4. Top 10 genes from the differential expression analysis in the combined Track-HD and Leiden cohort.

4.5.2 Pathway analysis

Genes were combined into networks with similar functional annotation, then expression was investigated in HD relative to controls. Pathway annotations were collated from publicly available gene ontology databases to form a set of generic pathways using the same method as the recent HD genome-wide association study (GWAS) of age at onset (GeM-HD, 2015, Hensman Moss et al., 2017a).

The number of pathways significantly dysregulated in both Track-HD and Leiden blood datasets was significantly higher than would be expected by chance. This indicates shared biology between the two independent cohorts despite differences in demographic and disease stage; Leiden subjects were on average 7 years older and had correspondingly higher TMS (mean 32 versus 14 in Track-HD) and lower TFC (mean 8 versus 12 in Track-HD). The significance of the overlap was greatly increased in analyses specifying the direction of dysregulation (increased or decreased expression) (see table below). Therefore, directional analyses were used in the combined dataset as the primary analysis.

Reference dataset	Comparison dataset	Direction of dysregulation in HD	Number of pathways significant in both datasets (p value)		
			Generic pathways	HD brain modules	Control brain modules
Leiden	Track-HD	Nondirectional	69 (4.6E-02)	-	-
		Downregulated	130 (<1.0E-03)	4 (1.1E-01)	24 (<1.0E-03)
		Upregulated	219 (<1.0E-03)	9 (<1.0E-03)	23 (<1.0E-03)
Track-HD	Leiden	Nondirectional	69 (1.4E-01)	-	-
		Downregulated	130 (1.7E-02)	4 (3.5E-02)	24 (<1.0E-03)
		Upregulated	217 (<1.0E-03)	10 (<1.0E-03)	21 (<1.0E-03)

Table 4.5. Overlap analysis of Track-HD and Leiden cohorts shows that a significant excess of pathways are associated with HD (p < 0.05) in both datasets.

Significance of overlap is greatest when directionality is taken into account. There is an excess of significantly enriched pathways and modules in the reference dataset conditional on the pathway being enriched (p < 0.05) in the comparison dataset. The generic pathways gene set is collated from publicly-available databases including GO and KEGG. HD brain modules are derived from Neueder and Bates (2014). Control brain modules are from the Braineac (2016) and Gibbs et al. (2010) expression datasets.

Gene set enrichment analysis (GSEA), with a false discovery rate (q-value) threshold of q < 0.05 to correct for multiple testing, identified 53 upregulated and 14 downregulated pathways that are at least nominally significant in both cohorts. Multiple immune-related pathways were upregulated, and RNA processing, ATP metabolism and DNA repair were notably downregulated. The 10 most significant pathways for each direction of dysregulation are given in in the table below and the full list of significant pathways is available in Hensman Moss et al. (2017a).

Direction of dysregulation in HD	Pathway	Number of dysregulated genes	p (combined)	q (combined)	p (Track-HD)	p (Leiden)	Description
Upregulated	MGI: 2419	434	3.03E-10	4.32E-06	5.10E-05	3.01E-05	Abnormal Innate Immunity
	MGI: 3009	432	5.78E-09	4.13E-05	5.96E-06	1.65E-04	Abnormal Cytokine Secretion
	GO: 50792	117	2.59E-08	1.23E-04	1.12E-02	7.24E-05	Regulation Of Viral Process
	GO: 9615	208	1.22E-07	4.36E-04	3.06E-02	5.34E-06	Response To Virus
	MGI: 2451	278	1.68E-07	4.80E-04	1.26E-02	9.51E-06	Abnormal Macrophage Physiology
	GO: 19221	308	2.38E-07	5.45E-04	4.60E-05	1.71E-04	Cytokine-Mediated Signaling Pathway
	GO: 2252	365	3.10E-07	5.45E-04	7.01E-03	1.14E-04	Immune Effector Process
	MGI: 5025	406	3.44E-07	5.45E-04	5.91E-05	2.02E-04	Abnormal Response To Infection
	MGI: 1793	372	4.33E-07	5.82E-04	5.93E-05	2.42E-04	Altered Susceptibility To Infection
	MGI: 8568	305	4.49E-07	5.82E-04	4.79E-05	6.25E-05	Abnormal Interleukin Secretion
Downregulated	GO: 8380	282	5.22E-08	7.45E-04	4.25E-05	7.24E-05	RNA splicing
	GO: 6397	359	2.38E-07	1.70E-03	1.48E-04	4.14E-04	mRNA processing
	GO: 16887	329	1.37E-06	5.48E-03	1.96E-04	3.34E-02	ATPase activity
	GO: 6200	333	1.54E-06	5.48E-03	2.42E-04	3.36E-02	ATP catabolic process
	GO: 46034	361	5.36E-06	1.53E-02	1.74E-04	4.45E-02	ATP metabolic process
	GO: 16607	144	9.06E-06	2.15E-02	4.68E-04	4.61E-03	Nuclear speck
	GO: 6281	356	1.66E-05	2.75E-02	2.00E-03	1.18E-04	DNA repair
	GO: 16604	271	2.08E-05	2.75E-02	5.59E-03	2.46E-03	Nuclear Body
	GO: 4386	135	2.12E-05	2.75E-02	2.83E-02	4.81E-02	Helicase Activity
	GO: 375	184	2.40E-05	2.86E-02	1.14E-03	2.05E-03	RNA splicing, via transesterification reactions

Table 4.6. The 10 most significantly up and downregulated 'generic' pathways in HD versus control blood GSEA.

A total of 14,706 Generic pathways, each containing between 3 and 500 genes, were collated from publicly-available databases including GO and KEGG. Pathways are significantly dysregulated after multiple testing correction (q < 0.05). Enrichment p values in the current study for the Track-HD, Leiden and combined datasets are shown.

Figure 4.1. Upregulated pathways in HD versus control blood.

Schematic representation of pathways collated from publicly available databases that are significantly upregulated in HD versus controls after correction for multiple testing (q < 0.05). Modules with similar gene content and functional annotation have been consolidated. Nodal shading is inversely proportional to false discovery rate threshold (q value); deep shades have low q values and pale shading is close to the 5% threshold. The weight of connecting lines is proportional to the number of genes shared between pathways.

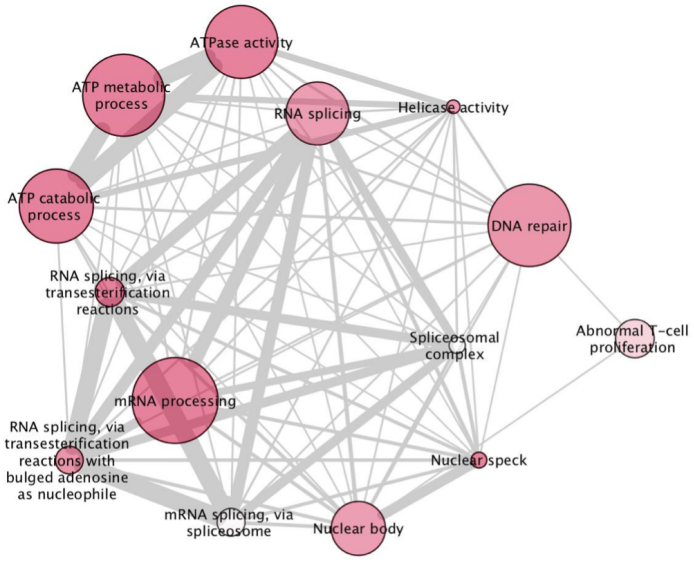


Figure 4.2. Downregulated pathways in HD versus control blood.

Schematic representation of pathways collated from publicly available databases that are significantly downregulated in HD versus controls after correction for multiple testing (q < 0.05). Modules with similar gene content and functional annotation have been consolidated. Nodal shading is inversely proportional to false discovery rate threshold (q value); deep shades have low q values and pale shading is close to the 5% threshold. The weight of connecting lines is proportional to the number of genes shared between pathways.

The 10 most dysregulated genes (p < 0.01) from the significantly up or downregulated generic pathways (q < 0.05) are shown below, and a complete list of genes (p < 0.05) in all nominally significant pathways (p < 0.05) is given in Hensman Moss et al. (2017a). Notably, the significantly upregulated pathways contain some of the most differentially expressed individual transcripts, with several more contained in pathways reaching nominal significance (p < 0.05) for dysregulation. Genes highlighted by MGI pathways appear distinct from other pathway databases, likely because they are based on knockout studies in mice.

Direction	Entrez gene ID	Gene Symbol	p (Comb)	log2FC (Comb)	p (Track-HD)	log2FC (Track-HD)	p (Leiden)	log2FC (Leiden)	Pathway membership (q < 0.05)
Genes in upregulated pathways	722	C4BPA	7.81E-06	1.371	1.29E-01	0.437	7.36E-01	0.187	GO:2252, GO:2253, GO:5773, GO:31347, GO:44437, GO:50778
	8763	CD164	9.53E-04	0.098	2.97E-01	0.083	5.57E-03	0.101	GO:323, GO:5764, GO:5765, GO:5773, GO:44437
	597	BCL2A1	1.06E-03	0.423	8.85E-02	0.319	1.20E-02	0.393	MGI:1793, MGI:2419, MGI:2462, MGI:2463, MGI:5025
	4940	OAS3	1.12E-03	0.688	5.14E-02	0.602	6.45E-02	0.455	GO:2252, GO:9615, GO:19221, GO:34340, GO:43903, GO:45069, GO:45071,GO:48525, GO:50792, GO:60337, GO:71345, GO:71357, KEGG:5164, REACTOME:287,REACTOME:587, REACTOME:589
	49	ACR	1.13E-03	1.237	7.54E-03	1.417	1.79E-01	0.768	GO:5773, GO:44437
	9262	STK17B	1.19E-03	0.132	4.42E-02	0.134	3.56E-02	0.136	MGI:1844, MGI:2425, MGI:2444, MGI:3009, MGI:5000, MGI:5005, MGI:8568
	164668	APOBEC3H	1.98E-03	-0.323	1.41E-01	-0.208	4.33E-03	-0.476	GO:2252, GO:9615, GO:43903, GO:45069, GO:45071, GO:48525, GO:50792
	79026	AHNAK	2.12E-03	-0.169	1.48E-02	-0.201	1.27E-01	-0.106	MGI:1793, MGI:2406, MGI:2444, MGI:3009, MGI:5025, MGI:8568
	6614	SIGLEC1	4.39E-03	0.634	3.58E-01	0.291	9.79E-02	0.552	MGI:2459, MGI:8195
	875	CBS	4.42E-03	0.592	1.15E-01	0.439	2.38E-02	0.681	MGI:8469, MGI:8713, MGI:8835
Genes in downregulated pathways	9262	STK17B	1.19E-03	0.132	4.42E-02	0.134	3.56E-02	0.136	MGI:5094
	54957	TXNL4B	1.65E-03	0.088	2.99E-02	0.088	2.67E-02	0.090	GO:5681, GO:6397, GO:8380
	375757	SWI5	1.68E-03	0.114	3.22E-02	0.112	2.67E-02	0.130	GO:6281
	146713	RBFOX3	1.86E-03	-0.434	3.81E-02	-0.396	7.65E-02	-0.357	GO:6397, GO:8380
	79026	AHNAK	2.12E-03	-0.169	1.48E-02	-0.201	1.27E-01	-0.106	MGI:5094
	29890	RBM15B	2.67E-03	-0.055	9.18E-02	-0.048	8.98E-02	-0.044	GO:6397, GO:8380
	9987	HNRNPDL	3.38E-03	-0.078	2.98E-02	-0.088	9.41E-03	-0.098	GO:5681
	23499	MACF1	3.72E-03	-0.120	4.52E-03	-0.172	2.15E-01	-0.068	GO:6200, GO:16887, GO:46034
	146754	DNAH2	4.04E-03	-0.621	1.82E-01	-0.415	2.39E-02	-0.723	GO:6200, GO:16887, GO:46034
	10236	HNRNPR	5.92E-03	-0.069	1.15E-01	-0.053	5.62E-02	-0.074	GO:375, GO:377, GO:398, GO:5681, GO:6397, GO:8380

Table 4.7. Top genes in top pathways.

The 10 most dysregulated genes (p < 0.01) from the significantly up or downregulated generic pathways (q < 0.05). p – corrected significance in the indicated dataset. Log2FC – Log2 fold change in expression.

Group		Pathway	Description	p (blood combined)
Group 1	Abnormal innate immunity	MGI: 2419	abnormal innate immunity	3.03E-10
		MGI: 2451	abnormal macrophage physiology	1.68E-07
		MGI: 2462	abnormal granulocyte physiology	4.09E-05
		MGI: 2463	abnormal neutrophil physiology	9.20E-05
		MGI: 2498	abnormal acute inflammation	1.01E-04
Group 2	Abnormal cytokine secretion	MGI: 3009	abnormal cytokine secretion	5.78E-09
		MGI: 8568	abnormal interleukin secretion	4.49E-07
		MGI: 8835	abnormal intercellular signaling peptide or protein level	8.31E-06
		MGI: 8713	abnormal cytokine level	1.35E-05
		MGI: 8469	abnormal protein level	2.68E-05
		MGI: 10210	abnormal circulating cytokine level	8.70E-05
		MGI: 8704	abnormal interleukin-6 secretion	1.54E-04
		MGI: 8705	increased interleukin-6 secretion	2.00E-04
Group 3	Regulation of viral process	GO: 50792	regulation of viral process	2.59E-08
		GO: 48525	negative regulation of viral process	6.05E-07
		GO: 45069	regulation of viral genome replication	1.21E-06
		GO: 45071	negative regulation of viral genome replication	2.30E-06
		GO: 43903	regulation of symbiosis, encompassing mutualism through parasitism	4.28E-05
		GO: 9615	response to virus	1.22E-07
		GO: 2252	immune effector process	3.10E-07
Group 4	Cytokine-mediated signalling pathway	GO: 19221	cytokine-mediated signaling pathway	2.38E-07
		REACTOME 287	REACT:CYTOKINE SIGNALING IN IMMUNE SYSTEM	8.59E-07
		GO: 71345	cellular response to cytokine stimulus	1.15E-06
		REACTOME 589	REACT:INTERFERON SIGNALING	3.02E-05
		GO: 60337	type I interferon-mediated signaling pathway	4.20E-05
		GO: 71357	cellular response to type I interferon	4.20E-05
		REACTOME 587	REACT:INTERFERON ALPHA BETA SIGNALING	6.00E-05
		GO: 34340	response to type I interferon	8.95E-05
		KEGG 5164	KEGG INFLUENZA A	1.14E-04
Group 5	Abnormal response to infection	MGI: 5025	abnormal response to infection	3.44E-07
		MGI: 1793	altered susceptibility to infection	4.33E-07
		MGI: 2406	increased susceptibility to infection	6.87E-06
Group 6	Abnormal myeloid leukocyte morphology	MGI: 8250	abnormal myeloid leukocyte morphology	6.46E-07
		MGI: 8251	abnormal phagocyte morphology	3.12E-06
		MGI: 8195	abnormal antigen presenting cell morphology	1.94E-05
		MGI: 2441	abnormal granulocyte morphology	3.42E-05
		MGI: 8248	abnormal mononuclear phagocyte morphology	4.39E-05
		MGI: 2446	abnormal macrophage morphology	5.88E-05
Group 7	Regulation of cytokine production	GO: 1817	regulation of cytokine production	2.76E-06
		GO: 31347	regulation of defense response	4.73E-06
		GO: 50778	positive regulation of immune response	7.29E-06
		GO: 2253	activation of immune response	2.92E-05
Group 8	Abnormal T-cell physiology	MGI: 2444	abnormal T cell physiology	1.14E-05
		MGI: 2459	abnormal B cell physiology	1.11E-04
Group 9	Abnormal self tolerance	MGI: 5005	abnormal self tolerance	1.21E-05
		MGI: 1844	autoimmune response	1.28E-05
		MGI: 5000	abnormal immune tolerance	1.63E-05
		MGI: 2425	altered susceptibility to autoimmune disorder	7.94E-05
Group 10	Vacuole	GO: 5773	vacuole	1.36E-05
		GO: 323	lytic vacuole	1.41E-05
		GO: 5764	lysosome	1.41E-05
		GO: 44437	vacuolar part	4.50E-05
		GO: 5765	lysosomal membrane	4.97E-05

Table 4.8. Groups of pathways upregulated in HD blood vs controls.

Group		Pathway	Description	p (blood-combined)
Group 1	mRNA splicing	GO: 8380	RNA splicing	5.22E-08
		GO: 6397	mRNA processing	2.38E-07
		GO: 375	RNA splicing, via transesterification reactions	2.40E-05
		GO: 377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	6.25E-05
		GO: 398	mRNA splicing, via spliceosome	6.25E-05
		GO: 5681	spliceosomal complex	7.29E-05
Group 2	ATPase activity	GO: 16887	ATPase activity	1.37E-06
		GO: 6200	ATP catabolic process	1.54E-06
		GO: 46034	ATP metabolic process	5.36E-06
Ungrouped terms		GO: 16607	nuclear speck	9.06E-06
		GO: 6281	DNA repair	1.66E-05
		GO: 16604	nuclear body	2.08E-05
		GO: 4386	helicase activity	2.12E-05
		MGI: 5094	abnormal T cell proliferation	4.60E-05

Table 4.9. Groups of pathways downregulated in HD blood vs controls.

4.5.3 Pathway dysregulation overlap with HD myeloid cells

With RNA-Seq, Miller et al. (2016) identified transcriptional dysregulation in unstimulated monocytes from HD cases relative to controls. Their GSEA used the same set of generic pathways used here. A significant excess of pathways were found to be significantly ($p<0.05$) enriched for dysregulation in both Miller et al. (2016) and the combined TRACK-HD and Leiden whole blood data. This overlap was attributable to a significant excess of pathways enriched for upregulation in both datasets. Overlap in downregulated pathways was not significantly larger than expected by chance. The top 10 pathways significantly ($p<0.05$) enriched for upregulation in both myeloid and whole blood are given in the table below, and the full list of up and downregulated pathways is listed in Hensman Moss et al. (2017a). Pathways that are significantly enriched for upregulation are predominantly immune-related, which is unsurprising given Miller et al. (2016) isolated monocytes from blood.

Direction of dysregulation in HD	Number of pathways significant in both datasets (p value)
Nondirectional	132 (0.009)
Downregulated	36 (0.113)
Upregulated	339 (<1.0E-03)

Table 4.10. Overlap between HD blood and myeloid cells.

A significant excess of pathways were found to be significantly ($p<0.05$) enriched for dysregulation in both Miller et al. (2016) and the combined TRACK-HD and Leiden whole blood data.

Pathway	Number of genes	p (blood: London+Leiden)	p (myeloid-unstimulated)	p (blood and myeloid combined)	Description
MGI: 2419	434	3.03E-10	3.77E-08	4.44E-16	abnormal innate immunity
MGI: 3009	432	5.78E-09	4.26E-07	8.54E-14	abnormal cytokine secretion
GO: 31347	430	4.73E-06	8.96E-09	1.35E-12	regulation of defense response
GO: 9615	208	1.22E-07	9.68E-07	3.64E-12	response to virus
MGI: 2451	278	1.68E-07	1.83E-06	9.16E-12	abnormal macrophage physiology
GO: 2252	365	3.10E-07	1.95E-06	1.76E-11	immune effector process
MGI: 1793	372	4.33E-07	2.30E-06	2.85E-11	altered susceptibility to infection
MGI: 8568	305	4.49E-07	3.26E-06	4.13E-11	abnormal interleukin secretion
MGI: 5025	406	3.44E-07	4.42E-06	4.28E-11	abnormal response to infection
MGI: 8835	258	8.31E-06	1.92E-07	4.49E-11	abnormal intercellular signaling peptide or protein level

Table 4.11. Top 10 upregulated pathways that overlap between HD blood and myeloid cells.
Pathways significantly (p<0.05) enriched for up and downregulation in both myeloid and whole blood.

4.5.4 Gene co-expression modules

4.5.4.1 HD blood

A limitation of using curated pathways from databases is the incomplete or incorrect annotation. One way to overcome this is to use gene co-expression, because genes that are co-expressed often have related functions. WGCNA identifies clusters (modules) of genes with highly correlated expression, constructing original, unbiased gene co-expression networks based on observed data (Gibbs et al., 2013). HD brain expression modules were generated by Neueder and Bates (2014), who applied WGCNA to Hodges et al. (2006) data and annotated each module that was associated with HD disease status. To further fill the annotation gap and better define functional biological pathways, novel co-expression modules were generated for control brain from the Braineac (2016) and Gibbs et al. (2010) datasets.

GSEA for brain co-expression modules was applied to the combined Track-HD and Leiden blood expression dataset. Immune and inflammatory-related brain modules were upregulated in HD blood, and notable downregulated modules included synaptic function, proteasomal degradation, mitochondrial function and transcription, as shown in the table below. Also given below is a table of the top 10 genes from the modules that are themselves nominally significantly dysregulated ($p < 0.05$) in the combined dataset, and the full list is provided in Hensman Moss et al. (2017a). In addition to reinforcing the biological conclusions from the previous pathway analysis, the significantly dysregulated modules also share genes with the top pathways above.

Direction	Brain expression gene set	Module	Brain region	Annotation	Number of genes	p (Combined)	p (Track-HD)	p (Leiden)	Cor (HD)	BH (HD)
Upregulated	HD	111	FC_BA9	Immune response	514	7.81E-12	1.27E-04	7.53E-05	-	-
	HD	69 (FC4pos1)	FC_BA4	Inflammatory response	712	3.77E-08	3.05E-05	1.32E-03	0.61	3.77E-03
	Control (B)	712	TCTX	Inflammatory response	213	1.41E-07	3.40E-05	8.14E-04	-	-
	HD*	**48 (CNpos2)***	**CN**	**Lipid metabolism/regulation of transcription**	**1785**	**2.03E-07**	**3.85E-03**	**6.33E-03**	**0.72**	**2.21E-11**
	Control (B)	110	FCTX	Inflammatory response	173	8.94E-07	1.04E-03	2.50E-03	-	-
	Control (B)	909	White Matter	Activation of immune response	265	2.12E-06	1.24E-03	2.48E-02	-	-
	Control (B)	610	Substantia Nigra	Inflammatory response	178	1.21E-05	8.56E-04	5.57E-04	-	-
	Control (B)	811	Thalamus	Inflammatory response	142	1.61E-05	3.94E-03	2.89E-03	-	-
	Control (G)	56	Pons	Lipoprotein/ immune response /GTPase regulator activity	207	1.97E-05	2.44E-04	4.19E-02	-	-
	Control (B)	911	White Matter	Inflammatory response	159	3.00E-05	8.42E-04	1.39E-02	-	-
	HD	28	CB	Immune response	209	3.11E-05	1.07E-02	1.19E-02	-	-
	Control (B)	713	TCTX	Activation of immune response	171	4.02E-05	2.39E-02	4.67E-02	-	-
	HD	33	CB	Immune response	255	4.34E-05	1.08E-02	1.37E-02	-	-
	Control (B)	505	Putamen	Ether lipid metabolism	500	6.28E-05	3.16E-03	2.06E-02	-	-
	HD	68 (CNpos5)	CN	Cilium	1268	1.09E-04	3.05E-02	5.00E-02	0.54	7.74E-06
	Control (B)	516	Putamen	Cellular response to cytokine stimulus	133	3.07E-04	1.44E-02	1.71E-02	-	-
	HD	64 (CNpos6)	CN	Inflammatory response	114	3.13E-04	1.18E-02	3.80E-02	0.46	2.28E-04
	HD	124	FC_BA9	NA	1176	2.91E-03	1.19E-02	2.37E-02	-	-
Downregulated	Control (G)	22	CB	Pro-rich region	831	1.83E-08	2.49E-03	2.06E-02	-	-
	Control (G)	28	FC	Intra-cellular transport/mitochondrion	3178	2.10E-08	6.30E-04	7.66E-05	-	-
	Control (B)	304	Medulla	mRNA metabolic process	1811	2.91E-08	5.00E-15	4.01E-02	-	-
	HD*	**66 (CNneg1)***	**CN**	**Synapse/ion channels**	**2645**	**2.71E-07**	**1.51E-04**	**2.13E-02**	**-0.80**	**6.03E-15**
	Control (B)	804	Thalamus	Regulation of cell morphogenesis	857	1.31E-06	4.03E-02	4.13E-04	-	-
	Control (B)	522	Putamen	Regulation of RNA splicing	64	4.44E-06	6.26E-03	2.66E-04	-	-
	Control (G)	74	Pons	Transcription/acetylation/protein transport	1183	9.22E-06	3.85E-08	7.44E-04	-	-
	Control (B)	702	TCTX	Antigen processing: ubiquitination and proteasome degradation	4602	3.87E-04	1.22E-03	2.47E-02	-	-
	Control (G)	48	FC	Transcription corepressor/cell morphogenesis	648	4.65E-04	7.83E-03	2.05E-02	-	-
	Control (B)	202	Hippocampus	Mitochondrial membrane	2737	4.75E-04	1.16E-07	1.54E-02	-	-
	HD	19	CB	Protein binding	155	7.44E-04	2.66E-02	2.26E-02	-	-
	Control (B)	906	White Matter	Uridyltransferase activity	416	1.12E-03	2.53E-02	1.12E-02	-	-
	Control (G)	93	Pons	Mitochondrion/nuclear lumen	317	1.30E-03	9.85E-03	8.74E-04	-	-
	Control (B)	812	Thalamus	Transport of mature transcript to cytoplasm	114	1.42E-03	1.99E-02	4.70E-02	-	-
	HD	102	FC_BA9	Cytoplasm	1908	1.47E-03	7.57E-03	1.31E-04	-	-
	Control (B)	706	TCTX	Microtubule organising center	481	1.93E-03	3.70E-05	3.80E-03	-	-
	Control (G)	52	Pons	Acetylation/fatty acid metabolism	1590	3.28E-03	2.23E-02	1.31E-02	-	-
	HD	3 (CBneg2)	CB	mitochondrion	1164	3.19E-02	2.56E-02	1.29E-05	-0.45	1.66E-03
	Control (G)	25	CB	RNA binding	648	8.02E-01	1.72E-04	3.62E-02	-	-

Table 4.12. WGCNA brain expression modules in HD versus control blood.

*P values for dysregulation in the combined blood sample are corrected for multiple testing ($q < 0.05$). HD brain modules were defined by Neueder and Bates (2014), and Control brain modules were generated from Braineac (2016) and Gibbs et al. (2010). Neueder and Bates (2014) module identifiers are given in brackets where available. * denotes the caudate modules that were highly positively or negatively correlated with HD in their study. BH – Benjamini-Hochberg correction for false discovery rate; CN – caudate nucleus; FC – frontal cortex; FC BA4 – BA4 region of the frontal cortex; FC BA9 – BA9 region of the frontal cortex; CB – cerebellum; TCTX – temporal cortex.*

Entrez gene ID	Gene Symbol	p (Comb)	log2FC (Comb)	p (Track-HD)	log2FC (Track-HD)	p (Leiden)	log2FC (Leiden)	Module membership
2297	FOXD1	9.09E-05	-0.785	1.10E-02	-0.685	1.69E-03	-1.014	HD 48 (CNpos2), HD 111
3805	KIR2DL4	1.93E-04	0.651	2.57E-03	0.823	1.52E-02	0.533	CTRL (B) 702
196394	AMN1	2.11E-04	0.208	1.87E-02	0.205	9.25E-03	0.195	CTRL (B) 202, CTRL (B) 702
5797	PTPRM	3.12E-04	-0.359	5.26E-03	-0.381	2.82E-03	-0.448	CTRL (B) 202, CTRL (B) 702, CTRL (B) 904, HD 66 (CNneg1)
889	KRIT1	7.30E-04	-0.081	1.59E-02	-0.097	7.86E-02	-0.057	CTRL (B) 304, CTRL (G) 28, HD 102
22979	EFR3B	8.17E-04	0.494	6.02E-03	0.603	2.07E-02	0.496	CTRL (B) 702, CTRL (B) 904, HD 66 (CNneg1)
56934	CA10	8.42E-04	2.036	1.21E-02	2.020	8.42E-02	1.945	CTRL (B) 702, CTRL (B) 902, HD 66 (CNneg1), HD 102
8763	CD164	9.53E-04	0.098	2.97E-01	0.083	5.57E-03	0.101	HD 68 (CNpos5)
597	BCL2A1	1.06E-03	0.423	8.85E-02	0.319	1.20E-02	0.393	CTRL (B) 110, CTRL (B) 217, CTRL (B) 516, CTRL (B) 610, CTRL (B) 712, CTRL (B) 811, CTRL (B) 911,HD 33, HD 68 (CNpos5), HD 69 (FC4pos1), HD 111
4940	OAS3	1.12E-03	0.688	5.14E-02	0.602	6.45E-02	0.455	CTRL (B) 702, CTRL (B) 902

Table 4.13. Top 10 genes in WGCNA modules.

Genes from the modules that are themselves nominally significantly dysregulated ($p < 0.05$) in the combined dataset.

4.5.4.2 Comparison of blood with HD striatum

Neueder and Bates (2014) derived 124 HD brain expression modules in four brain regions by applying WGCNA to the Hodges et al. (2006) microarray expression dataset of 44 human HD and 36 matched control brains. The table below gives modules that were significantly dysregulated (after correcting for multiple testing of modules) in both HD brain (Neueder and Bates, 2014) and in the combined Track-HD and Leiden blood expression dataset. The direction of dysregulation in brain is shown by the correlation between the module eigengene and HD status (with a positive correlation corresponding to upregulation in the HD brain). Notably, two of the most significantly dysregulated modules in HD caudate (Neueder and Bates, 2014) were also significantly dysregulated in the same direction in blood, not only in the combined dataset, but in each of the Track-HD and Leiden datasets independently; these being module 48 (CNpos2), which is upregulated in HD, and module 66 (CNneg1), which is downregulated.

Module	Brain Region	Module name	Number of genes	p (combined)	p (TRACK)	p (Leiden)	cor (HD brain)	p (HD brain)	Description
69	FC_BA4	FC4pos1	712	3.77E-08	3.05E-05	1.32E-03	0.610	3.77E-03	Inflammatory response
48	CN	CNpos2	1785	2.03E-07	3.85E-03	6.33E-03	0.724	2.21E-11	Lipid metabolism/regulation of transcription
64	CN	CNpos6	114	3.13E-04	1.18E-02	3.80E-02	0.463	2.28E-04	Inflammatory response
66	CN	CNneg1	2644	2.71E-07	1.51E-04	2.13E-02	-0.800	6.03E-15	Synapse

Table 4.14. Brain expression modules significantly dysregulated both in HD brain and HD blood.
All modules in this table are significantly dysregulated after correction for multiple testing (q < 0.05) in the combined blood sample, and are nominally significantly dysregulated (p<0.05) in both Track-HD and Leiden datasets separately. Cor(HD brain) – the correlation between module eigengene and HD status observed by Neueder and Bates (2014) in brain expression data, with a positive correlation corresponding to upregulation in HD. p(HD brain) is the p-value for that correlation (corrected for multiple testing of modules). CN – caudate nucleus, FC_BA4 – BA4 region of the frontal cortex.

The module membership (kME) of a gene is measured by the correlation of its expression with the eigengene, which is representative of all gene expression profiles in the module (Langfelder and Horvath, 2008); highly connected 'hub' genes have high kME values. Interestingly, among genes in module 48 (CNpos2), the Neueder and Bates (2014) HD caudate module that was also significantly upregulated in blood, there was a significant ($p = 7.6 \times 10^{-4}$) correlation between dysregulation p-value in the direction of interest (positive) in HD blood and degree of module membership (kME) (Neueder and Bates, 2014). This suggests that highly connected "hub" genes in this module may play a role in transcriptional dysregulation in HD. A similar, although much stronger, effect was noted in caudate (Neueder and Bates, 2014). There was no significant correlation in module 66 (CNneg1). The top 10 genes in module 48 (CNpos2) that are dysregulated ($p < 0.05$) in both blood and caudate are shown in below, ranked by their kME value, and the full list is given in Hensman Moss et al. (2017a).

Gene	Entrez	kME	log2FC (blood)	Directional p (blood)	log2FC (caudate)	p (caudate)
RAB13	5872	0.877	0.138	3.11E-02	0.761	8.07E-10
RAB31	11031	0.858	0.118	4.82E-02	0.503	6.25E-09
MT2A	4502	0.833	0.237	1.40E-02	0.640	1.98E-06
S100A6	6277	0.810	0.121	2.37E-02	0.776	3.13E-09
DDIT4	54541	0.781	0.193	9.64E-03	0.827	4.82E-05
PFKFB3	5209	0.780	0.286	5.78E-03	0.568	1.39E-06
TSPO	706	0.772	0.170	3.30E-03	0.452	6.60E-08
DDAH2	23564	0.762	0.087	4.48E-02	0.375	4.04E-09
ITGA7	3679	0.758	0.151	4.38E-02	0.287	1.21E-04
PLEKHF2	79666	0.742	0.067	3.34E-02	0.370	1.16E-06

Table 4.15. Top 10 genes in module 48 (CNpos2) that are dysregulated (p < 0.05) in both blood and caudate, ranked by their kME value.
Module membership (kME) of a gene is measured by the correlation of its expression with the eigengene, which is representative of all gene expression profiles in the module; highly connected 'hub' genes have high kME values.

4.5.4.3 Comparison of blood with HD prefrontal cortex

Labadorf et al. (2015) identified dysregulated expression of immune and developmental genes in human HD postmortem prefrontal cortex (Brodmann area 9). Fold changes in expression of individual genes in the combined Track-HD and Leiden data were compared to those observed in Labadorf et al. (2015), and were found to be in the same direction for 8,425 out of the 15,834 genes present in both datasets. This is a highly significant ($p < 2.2 \times 10^{-16}$) excess (see Methods), suggesting some concordance in signal at the individual gene level.

Furthermore, a significant excess of generic pathways was found to be significantly ($p < 0.05$) dysregulated in both datasets, most markedly in the positive ($p < 0.001$) direction, but also negative ($p = 0.028$), thus showing an overlap in biological signal. The top 10 pathways up and downregulated in HD blood and prefrontal cortex are given in the table below, and the full list is available in Hensman Moss et al. (2017a). Pathways significantly upregulated in both datasets are mainly related to immune response.

Direction	Pathway	Number of dysregulated genes	Blood p (Combined)	Brain p (Labadorf)	Description
Upregulated	MGI: 2459	402	1.11E-04	1.39E-13	abnormal B cell physiology
	MGI: 2419	434	3.03E-10	2.05E-12	abnormal innate immunity
	MGI: 1800	361	5.45E-04	2.58E-12	abnormal humoral immune response
	MGI: 8195	412	1.94E-05	8.78E-12	abnormal antigen presenting cell morphology
	MGI: 2490	333	8.00E-04	3.52E-11	abnormal immunoglobulin level
	MGI: 8250	462	6.46E-07	4.04E-11	abnormal myeloid leukocyte morphology
	MGI: 4939	381	3.31E-03	1.68E-10	abnormal B cell morphology
	GO: 50778	403	7.29E-06	2.11E-10	positive regulation of immune response
	MGI: 8251	387	3.12E-06	3.29E-10	abnormal phagocyte morphology
	MGI: 3009	432	5.78E-09	5.24E-10	abnormal cytokine secretion
Downregulated	GO: 5874	327	4.97E-05	8.10E-05	microtubule
	GO: 86	120	8.11E-03	1.70E-04	G2/M transition of mitotic cell cycle
	GO: 48812	455	3.45E-02	2.20E-04	neuron projection morphogenesis
	PAN-PW 29	120	4.80E-02	2.67E-04	Huntington disease
	GO: 15631	187	4.14E-04	2.84E-04	tubulin binding
	GO: 7017	372	7.94E-04	4.13E-04	microtubule-based process
	MGI: 1828	233	1.66E-04	7.14E-04	abnormal T cell activation
	REACTOME 214	68	3.16E-03	1.12E-03	REACT:centrosome maturation
	REACTOME 952	68	3.16E-03	1.12E-03	REACT:recruitment of mitotic centrosome proteins and complexes
	REACTOME 636	59	1.87E-03	1.48E-03	REACT:loss of nlp from mitotic centrosomes

Table 4.16. Top 10 pathways dysregulated ($p < 0.05$) in both HD prefrontal cortex (Labadorf et al., 2015) and blood.

The pattern of immune upregulation is also observed in the brain co-expression modules. The top 10 modules in HD prefrontal cortex and blood are given in the table below, and the full list is available in Hensman Moss et al. (2017a). Notably, several modules related to the synapse and neuron projection are downregulated in both datasets. The two HD-related caudate modules from Neueder and Bates (2014) that were significantly dysregulated in blood were also significantly dysregulated in the same direction in Labadorf et al. (2015). Module 48 (CNpos2), which enriches for transcriptional regulators, was significantly upregulated ($p < 1x10^{-16}$) and module 66 (CNneg1), enriched for synaptic genes, was significantly downregulated ($p < 1x10^{-16}$), as are several other significant modules from Neueder and Bates (2014).

Direction	Brain expression gene set	Module	Brain Region	Number of dysregulated genes	Blood p (combined)	Brain p (Labadorf)	Cor(HD)	p (HD)	Annotation
Upregulated	HD	111	FC BA9	514	7.81E-12	0.00E+00	-	-	Immune response
	HD	69 (FC4pos1)	FC BA4	712	3.77E-08	0.00E+00	0.610	3.77E-03	inflammatory response
	HD	**48 (CNpos2)**	CN	1785	2.03E-07	0.00E+00	0.724	2.21E-11	regulation of transcription
	Control (Braineac)	803	Thalamus	870	1.18E-03	0.00E+00	-	-	Inflammatory response
	HD	124	FC BA9	1176	2.91E-03	0.00E+00	-	-	protein binding
	HD	4 (CBpos6)	CB	781	6.17E-03	0.00E+00	0.309	4.79E-02	metallothionein
	Control (Braineac)	705	Temporal Cortex	494	1.09E-02	1.17E-15	-	-	Inflammatory response
	Control (Braineac)	908	White Matter	359	3.48E-02	2.55E-15	-	-	Activation of immune response
	Control (Gibbs)	49	Frontal Cortex	446	6.83E-03	3.50E-15	-	-	oxidoreductase
	Control (Braineac)	106	Frontal Cortex	430	2.53E-03	3.94E-15	-	-	activation of immune response
Downregulated	HD	**66 (CNneg1)**	CN	2644	2.71E-07	0.00E+00	-0.8	6.03E-15	synapse
	Control (Braineac)	602	Substantia Nigra	2374	1.07E-03	0.00E+00	-	-	mitochondrial envelope
	Control (Braineac)	103	Frontal Cortex	855	4.01E-02	0.00E+00	-	-	Neuron projection morphogenesis
	HD	98 (FC4neg5)	FC BA4	1359	1.97E-04	5.55E-17	-0.445	0.03354	glycolysis
	Control (Braineac)	802	Thalamus	3342	5.39E-03	1.11E-16	-	-	Neuron projection
	Control (Braineac)	4	Cerebellar Cortex	973	2.26E-05	1.14E-13	-	-	neuron projection
	Control (Braineac)	407	Occipital Cortex	432	1.10E-02	6.71E-12	-	-	Transmission across Chemical Synapses
	Control (Braineac)	302	Medulla	1911	2.63E-05	1.69E-11	-	-	Cellular respiration
	Control (Braineac)	708	Temporal Cortex	414	6.81E-03	3.86E-11	-	-	Autophagy
	HD	2 (CBneg4)	CB	408	1.22E-02	3.44E-10	-0.388	0.01111	mitochondrion

Table 4.17. Top 10 modules dysregulated (p < 0.05) in both HD prefrontal cortex (Labadorf et al., 2015) and blood.

4.5.5 Association with disease severity

4.5.5.1 Individual transcripts

To look for an effect on disease severity, gene expression was correlated with UHDRS total motor score (TMS). After correcting for multiple testing, expression of phosphatidylcholine transfer protein (PCTP) was significantly positively correlated with TMS. However, this gene was not found to be significantly correlated with TMS by Mastrokolias et al (Mastrokolias et al., 2015).

Entrez gene ID	Gene Symbol	p (corr-TMS)	q (corr-TMS)	log2(FC)
58488	PCTP	1.82E-06	3.25E-02	8.00E-03
51060	TXNDC12	4.42E-05	1.79E-01	5.30E-03
57096	RPGRIP1	4.64E-05	1.79E-01	1.25E-02
9258	MFHAS1	5.12E-05	1.79E-01	-8.40E-03
3667	IRS1	6.73E-05	1.79E-01	-1.37E-02
158293	FAM120AOS	6.88E-05	1.79E-01	3.80E-03
84263	HSDL2	7.01E-05	1.79E-01	6.30E-03
56925	LXN	1.01E-04	2.22E-01	1.05E-02
118881	COMTD1	1.12E-04	2.22E-01	-8.10E-03
597	BCL2A1	1.44E-04	2.35E-01	1.44E-02

Table 4.18. Top 10 genes with expression in correlation with disease severity (total motor score).

4.5.5.2 Gene sets

Generic pathways that were significantly enriched for up or downregulated genes (Table 4.6), also enriched for genes correlated with TMS in the expected direction using a similar method to that previously used to test for enrichment of differentially expressed genes. The top 10 up and downregulated pathways are given in the table below and the full list is in Hensman Moss et al. (2017a). Several immune related pathways were positively correlated with TMS, including MGI:2419, the most significantly dysregulated pathway in HD blood (Table 6). Downregulated pathways that correlated with TMS were related to ATP metabolism and DNA repair.

Direction	Pathway	p (combined-diffexp)	p (TRACK-diffexp)	p (TRACK-TMS)	Description
Upregulated	MGI: 2419	3.03E-10	5.10E-05	2.18E-03	abnormal_innate_immunity
	GO: 10942	8.79E-02	4.70E-02	3.21E-03	positive regulation of cell death
	MGI: 2462	4.09E-05	6.48E-04	6.39E-03	abnormal_granulocyte_physiology
	MGI: 8556	6.85E-04	8.68E-03	7.91E-03	abnormal_tumor_necrosis_factor_secretion
	MGI: 2463	9.20E-05	2.79E-03	8.99E-03	abnormal_neutrophil_physiology
	MGI: 8704	1.54E-04	4.76E-03	9.56E-03	abnormal_interleukin-6_secretion
	GO: 5773	1.36E-05	7.03E-03	1.62E-02	vacuole
	GO: 50792	2.59E-08	1.12E-02	1.64E-02	regulation of viral process
	MGI: 5351	6.95E-03	2.20E-02	2.76E-02	decreased_susceptibility_to_autoimmune_disorder
	GO: 44437	4.50E-05	6.10E-04	3.48E-02	vacuolar part
Downregulated	GO: 45786	8.92E-04	1.88E-02	3.23E-05	negative regulation of cell cycle
	MGI: 706	9.70E-02	1.11E-02	9.09E-05	small_thymus
	MGI: 2364	6.81E-02	1.14E-02	2.57E-04	abnormal_thymus_size
	MGI: 5018	6.50E-04	5.94E-03	2.70E-04	decreased_T_cell_number
	MGI: 2435	1.95E-04	6.87E-03	2.79E-04	abnormal_effector_T_cell_morphology
	MGI: 8081	1.48E-03	5.61E-03	3.83E-04	abnormal_single-positive_T_cell_number
	MGI: 2145	1.19E-03	5.67E-03	8.68E-04	abnormal_T_cell_differentiation
	MGI: 2444	3.61E-04	7.95E-03	8.74E-04	abnormal_T_cell_physiology
	MGI: 2432	6.45E-04	3.48E-02	1.01E-03	abnormal_CD4-positive_T_cell_morphology
	MGI: 6387	8.81E-05	4.95E-03	1.31E-03	abnormal_T_cell_number

Table 4.19. Top 10 pathways enriched for up and downregulation in HD blood that also enriched for genes correlated with disease severity (TMS) in the same direction.

4.5.5.3 Co-expression modules

Similarly, modules dysregulated in HD blood relative to controls (Table 4.12) were also correlated with TMS in the expected direction. As shown in the table below, many modules significantly correlated with TMS, including 68 (CNpos5; $p=5.52x10^{-7}$) and 66 (CNneg1; $p=1.05x10^{-7}$), which were also dysregulated in the HD caudate (Neueder and Bates, 2014).

Direction	Brain expression gene set	Module	Brain region	Annotation	Number of dysregulated genes	p (Combined-diffexp)	p (TRACK-diffexp)	p (TRACK-TMS)	Cor (HD)	BH (HD)
Upregulated	HD	68 (CNpos5)	CN	Cilium	1268	1.09E-04	3.05E-02	5.52E-07	0.54	7.74E-06
	Control (B)	909	White Matter	Activation of immune response	265	2.12E-06	1.24E-03	8.22E-04	-	-
	Control (B)	713	TCTX	Activation of immune response	171	4.02E-05	2.39E-02	1.69E-03	-	-
	HD	111	FC_BA9	Immune response	514	7.81E-12	1.27E-04	3.75E-03	-	-
	Control (G)	56	Pons	Lipoprotein/ immune response /GTPase regulator activity	207	1.97E-05	2.44E-04	7.72E-03	-	-
	HD	28	CB	Immune response	209	3.11E-05	1.07E-02	8.70E-03	-	-
	Control (B)	505	Putamen	Ether lipid metabolism	500	6.28E-05	3.16E-03	6.43E-02	-	-
	Control (B)	911	White Matter	Inflammatory response	159	3.00E-05	8.42E-04	7.75E-02	-	-
	HD	124	FC_BA9	NA	1176	2.91E-03	1.19E-02	9.14E-02	-	-
	Control (B)	110	FCTX	Inflammatory response	173	8.94E-07	1.04E-03	1.34E-01	-	-
	HD	33	CB	Immune response	255	4.34E-05	1.08E-02	1.52E-01	-	-
	Control (B)	610	Substantia Nigra	Inflammatory response	178	1.21E-05	8.56E-04	2.00E-01	-	-
	HD	64 (CNpos6)	CN	Inflammatory response	114	3.13E-04	1.18E-02	2.22E-01	0.46	2.28E-04
	Control (B)	811	Thalamus	Inflammatory response	142	1.61E-05	3.94E-03	2.28E-01	-	-
	Control (B)	712	TCTX	Inflammatory response	213	1.41E-07	3.40E-05	2.35E-01	-	-
	Control (B)	516	Putamen	Cellular response to cytokine stimulus	133	3.07E-04	1.44E-02	4.16E-01	-	-
	HD	69 (FC4pos1)	FC_BA4	Inflammatory response	712	3.77E-08	3.05E-05	5.22E-01	0.61	3.77E-03
	HD*	48 (CNpos2)	CN	Lipid metabolism/regulation of transcription	1785	2.03E-07	3.85E-03	6.14E-01	0.72	2.21E-11
Downregulated	Control (B)	304	Medulla	mRNA metabolic process	1811	2.91E-08	5.00E-15	6.11E-16	-	-
	Control (B)	702	TCTX	Antigen processing: ubiquitination and proteasome degradation	4602	3.87E-04	1.22E-03	2.04E-13	-	-
	Control (B)	202	Hippocampus	Mitochondrial membrane	2737	4.75E-04	1.16E-07	1.44E-09	-	-
	Control (G)	28	FC	Intra-cellular transport/mitochondrion	3178	2.10E-08	6.30E-04	4.16E-09	-	-
	HD*	66 (CNneg1)	CN	Synapse/ion channels	2645	2.71E-07	1.51E-04	1.05E-07	-0.80	6.03E-15
	Control (G)	52	Pons	Acetylation/fatty acid metabolism	1590	3.28E-03	2.23E-02	1.30E-07	-	-
	Control (G)	74	Pons	Transcription/acetylation/protein transport	1183	9.22E-06	3.85E-08	1.19E-05	-	-
	Control (G)	22	CB	Pro-rich region	831	1.83E-08	2.49E-03	7.72E-05	-	-
	Control (B)	804	Thalamus	Regulation of cell morphogenesis	857	1.31E-06	4.03E-02	8.29E-05	-	-
	Control (B)	706	TCTX	Microtubule organising center	481	1.93E-03	3.70E-05	3.00E-04	-	-
	Control (G)	48	FC	Transcription corepressor/cell morphogenesis	648	4.65E-04	7.83E-03	7.14E-04	-	-
	HD	102	FC_BA9	Cytoplasm	1908	1.47E-03	7.57E-03	9.26E-03	-	-
	Control (B)	906	White Matter	Uridyltransferase activity	416	1.12E-03	2.53E-02	1.34E-02	-	-
	Control (B)	812	Thalamus	Transport of mature transcript to cytoplasm	114	1.42E-03	1.99E-02	1.36E-02	-	-
	HD	19	CB	Protein binding	155	7.44E-04	2.66E-02	2.18E-02	-	-
	HD	3 (CBneg2)	CB	mitochondrion	1164	3.19E-02	2.56E-02	6.17E-02	-0.45	1.66E-03
	Control (G)	93	Pons	Mitochondrion/nuclear lumen	317	1.30E-03	9.85E-03	1.24E-01	-	-
	Control (B)	522	Putamen	Regulation of RNA splicing	64	4.44E-06	6.26E-03	2.52E-01	-	-
	Control (G)	25	CB	RNA binding	648	8.02E-01	1.72E-04	9.99E-01	-	-

Table 4.20. Modules dysregulated in HD blood that also correlated with disease severity (TMS) in the same direction.
Control (B) – co-expression modules generated from Braineac (2016), Control (G) – co-expression modules generated from Gibbs et al. (2010).

4.5.5.4 Overlap with an independent HD cohort

Mastrokolias et al. (2015) listed 170 genes significantly associated with TMS, of which 142 passed quality control in our RNA-Seq data. These were tested for correlation between TMS in gene positive subjects from the Track-HD cohort. The top 10 genes in Track-HD are given below and the full list is available from Hensman Moss et al. (2017a). 14 genes were nominally significant ($p < 0.05$), which is significantly higher than expected by chance ($p = 7.89 \times 10^{-3}$). Using the same method as for concordance with Labadorf et al. (2015) (see Methods), fold changes in expression of individual genes were compared between Track-HD and Mastrokolias et al (Mastrokolias et al., 2015). Strikingly, 101 genes showed consistent direction of effect, as measured by log(FC), significantly greater than expected by chance ($p=4.78 \times 10^{-7}$). Thus, the analysis of TMS in the Track-HD cohort broadly supports the associations reported in Mastrokolias et al. (2015).

Ensembl gene ID	Entrez gene ID	Gene name	log(FC)-Mastrokolias	p (Mastrokolias)	log(FC)-TRACK	p (TRACK)
ENSG00000119471	84263	HSDL2	7.00E-03	4.86E-02	6.00E-03	7.01E-05
ENSG00000110422	10114	HIPK3	7.00E-03	3.77E-02	-6.00E-03	9.52E-03
ENSG00000177542	79751	SLC25A22	-7.00E-03	3.58E-02	-3.00E-03	1.14E-02
ENSG00000103569	366	AQP9	1.20E-02	4.68E-02	7.00E-03	1.41E-02
ENSG00000185803	79581	SLC52A2	-8.00E-03	4.60E-02	-3.00E-03	1.50E-02
ENSG00000188322	388228	SBK1	-7.00E-03	4.54E-02	-5.00E-03	1.56E-02
ENSG00000101096	4773	NFATC2	-1.10E-02	1.69E-02	-8.00E-03	1.83E-02
ENSG00000171051	2357	FPR1	9.00E-03	2.77E-02	6.00E-03	2.39E-02
ENSG00000112159	23195	MDN1	-6.00E-03	9.10E-03	-5.00E-03	2.45E-02
ENSG00000008869	54497	HEATR5B	-7.00E-03	4.68E-02	-3.00E-03	2.69E-02

Table 4.21. Top 10 differentially expressed genes from Mastrokolias et al (Mastrokolias et al., 2015) that correlated with disease severity (TMS) in Track-HD blood.

4.5.6 Comparing the HD transcriptomic signature with Alzheimer's disease brain

In Alzheimer's disease, an early inflammatory response involving microglia contributes to pathogenesis (Gomez-Nicola et al., 2013, Olmos-Alonso et al., 2016, Hong et al., 2016a). Given the upregulation of immune-related gene sets in HD, co-expression modules dysregulated in Alzheimer's disease (AD) brain were tested to see if they are also disrupted in HD blood. The International Genomics of Alzheimer's Disease Consortium (IGAP) identified four modules from the Gibbs et al. (2010) brain co-expression network that showed enrichment of signal in the GWAS of >70,000 late-onset Alzheimer's disease (LOAD) and control subjects (International Genomics of Alzheimer's Disease, 2015). These four modules, each derived from a different brain region, are all involved in the immune response and were all significantly upregulated in the combined HD blood dataset; they are given in the table below. Module 56, derived from pontine data, was also significantly enriched in both Track-HD and Leiden datasets independently. IGAP identified 151 genes that were present in two or more of these modules and showed the most significant enrichment with LOAD GWAS signal (International Genomics of Alzheimer's Disease, 2015). These 151 genes were also significantly enriched for upregulation in the combined HD blood dataset ($p = 2.50 \times 10^{-4}$).

Module	Brain Region	Number of genes	p (IGAP)	p (Comb)	p (Track-HD)	p (Leiden)	Module Description
34	Frontal Cortex	109	1.00E-05	1.45E-03	7.06E-03	9.48E-02	GO:0006955 immune response
99	Temporal Cortex	145	4.00E-05	2.22E-04	5.25E-03	9.13E-02	GO:0006955 immune response
56	Pons	207	6.00E-05	1.97E-05	2.44E-04	4.19E-02	GO:0006955 immune response
5	Cerebellum	135	6.80E-04	1.09E-03	4.24E-02	8.15E-02	GO:0006955 immune response

Table 4.22. Modules from Gibbs et al. (2010) that are dysregulated in both Alzheimer's disease brain (International Genomics of Alzheimer's Disease, 2015) and HD blood.
Comb – combined TRACK-HD and Leiden bHD blood dataset.

Zhang et al. (2013) identified co-expression modules that were differentially connected between LOAD and controls. Ten of these were also significantly enriched for upregulation in the HD blood expression dataset (given in the table below) after correction for multiple testing ($q < 0.05$), with their most significant module, *yellow*, being particularly highly enriched (combined Track-HD and Leiden $p < 1 \times 10^{-16}$). Notably, this module has immune and microglia-specific functions (Zhang et al., 2013). This enrichment for modules from the IGAP GWAS (International Genomics of Alzheimer's Disease, 2015) and Zhang et al. (2013) in the HD blood transcriptome suggests a shared immune-related mechanism between different neurodegenerative diseases, at least including HD and Alzheimer's disease.

Module	Rank (Zhang)	Annotation	Brain region	Number of genes	p (Comb)	q (Comb)	p (Track-HD)	p (Leiden)
Yellow	1	Immune functions	PFC	867	<1.00E-16	3.32E-15	1.55E-11	4.93E-11
Cyan	5	Vasculature development	VC	487	2.71E-11	4.49E-10	4.65E-08	1.33E-05
Gold	25	Immune functions	CB	318	1.09E-10	1.21E-09	1.28E-04	1.85E-04
Light cyan	11	Immune functions	VC	434	2.14E-09	1.77E-08	1.95E-05	5.36E-04
Forestgreen	24	Immune functions	CB	200	9.86E-06	6.54E-05	3.23E-05	1.12E-02
Red	9	Nerve ensheathment (myelination)	PFC	701	2.04E-04	1.13E-03	4.50E-02	8.89E-03
Violet red	21	Nitric oxide and human cancer	VC	124	3.14E-04	1.49E-03	1.10E-02	8.24E-02
Navy	15	Regulation of cell growth	CB	200	1.55E-03	6.44E-03	3.68E-02	5.56E-03
Turquoise	13	NAD(P) homeostasis	VC	909	6.30E-03	2.32E-02	5.87E-02	4.38E-03
Green yellow	8	Unfolded protein	PFC	478	1.02E-02	3.38E-02	7.42E-02	3.67E-01

Table 4.23. Top 10 co-expression modules from Alzheimer's disease brain(Zhang et al., 2013) that are dysregulated in HD blood. *Significance was corrected for multiple testing (q < 0.05). PFC – prefrontal cortex, VC – visual cortex, CB – cerebellum.*

4.6 Discussion

HD research has focused on the brain because the most conspicuous clinical features are clearly linked to progressive degeneration of specific brain regions (van der Burg et al., 2009, Bates et al., 2015c). However, HD is a systemic condition with peripheral expression of mutant huntingtin directly driving abnormalities such as immune dysfunction, metabolic derangement and transcriptional dysregulation that contribute to onset, progression, quality of life and mortality (van der Burg et al., 2009, Carroll et al., 2015).

In this chapter, RNA-Seq of whole blood was conducted in two independent cohorts of HD patients. Using gene set enrichment analysis (GSEA) with publicly-available pathway databases and WGCNA modules from HD and control brain datasets, gene sets were found to be dysregulated in blood that replicated in both independent cohorts and correlated with clinical motor signs (TMS). These correspond to the most significantly dysregulated modules in caudate nucleus, the most prominently affected region in HD brain. This suggests mutant huntingtin drives a pathogenic signature that is common to both blood and brain.

4.6.1 Individual transcripts

RNA-Seq more comprehensively and accurately quantifies mRNA than hybridisation-based microarrays or tag-based methods (Costa et al., 2010). Expression of phosphatidylcholine transfer protein (PCTP) significantly correlated with TMS (Table 4.18). This protein transports phospholipids across intracellular membranes, which is of interest given the upregulation of modules representing lipid metabolism in both HD blood (Table 4.12) and brain (Neueder and Bates, 2014). Phospholipid levels are altered in in mouse models and human postmortem HD brain, HTT carrying an expanded polyglutamine tract can disrupt lipid bilayers, and HTT phospholipid binding is altered in HD, which may be involved in its aggregation (Kegel-Gleason, 2013). However, PCTP was not significantly correlated with TMS in Mastrokolias et al. (2015).

It is perhaps unsurprising that there was limited differential expression of individual transcripts by disease state (Table 4) or severity in either the independent or combined cohorts; the major cell types known to contribute to HD symptoms are not present in blood and the haematogenous cells known to be dysfunctional in HD, such as monocytes and macrophages (Bjorkqvist et al., 2008, Wild et al., 2011), constitute only a small proportion of circulating cells (Whitney et al., 2003). There is considerable variation of gene expression in blood with age, gender, cell type and time of day, which also likely limited sensitivity (Whitney et al., 2003, Horvath et al., 2012). Our results are consistent with previous studies that have shown weak correlation at the transcript level between blood and brain (Cai et al., 2010).

4.6.2 Gene sets

4.6.2.1 Immune upregulation

Despite these limitations, gene set enrichment analysis identified significantly overlapping dysregulated pathways in the Track-HD and Leiden HD blood datasets, even though they differed in age and disease severity. Therefore, through grouping transcripts into biologically relevant pathways and co-expressed transcripts, it was possible to highlight areas of dysfunctional biology in HD. The observed upregulation of immune-related pathways (Table 4.6) and modules (Table 4.12) is consistent previous transcriptional and functional studies (Mastrokolias et al., 2015, Carroll et al., 2015, van der Burg et al., 2009). HD patients are known to have immune dysfunction, both in the central nervous system (CNS) with microglial activation (Tai et al., 2007a), and peripherally with elevated proinflammatory cytokines in premanifest carriers

up to 16 years before predicted onset (Bjorkqvist et al., 2008, Wild et al., 2011). The migration of phagocytic cells is impaired in HD (Kwan et al., 2012c, Träger et al., 2015) and patient-derived monocytes are hyperactive on stimulation, an effect reduced by HTT lowering (Bjorkqvist et al., 2008). Modulation of the peripheral immune system with a type 2 cannabinoid receptor (CB2) agonist (Bouchard et al., 2012b) or bone marrow transplantation (Kwan et al., 2012a) can increase lifespan and reduce motor deficits and synaptic loss in HD mouse models.

4.6.2.2 RNA processing

RNA processing modules (Table 4.12) were downregulated, which is consistent with disruption of splicing, miRNA expression and processing (Seredenina and Luthi-Carter, 2012) in HD brain by dysregulation of splicing factors such as PTBP1 (Lin et al., 2016), and the disruption of nuclear mRNA export in mouse models (Gasset-Rosa et al., 2017).

4.6.2.3 Energy metabolism

Pathways (Table 4.6) and modules (Table 4.12) involved in mitochondrial function and energy metabolism were downregulated in HD blood, which is consistent with prominent deficits in energy metabolism, particularly mitochondrial function, seen in HD patients and animal models. In patients, glucose consumption is reduced and there is ATP depletion, particularly in the basal ganglia but also throughout the body, even in presymptomatic carriers, and the lactate-pyruvate ratio in CSF is elevated (Acuña et al., 2013, Mochel and Haller, 2011, Jodeiri Farshbaf and Ghaedi, 2017). Proposed mechanisms include impaired oxidative phosphorylation due to respiratory chain deficiencies, increased oxidative stress, as evidenced by increased reactive oxygen species, oxidative DNA damage and the induction of oxidative defence mechanisms in HD brain, and impaired trafficking and biogenesis of mitochondria. *PGC-1α*, a member of the downregulated *ATP metabolic process* pathway (Table 4.6), is a key protective regulator of mitochondrial genes and biogenesis. It has previously been shown to be reduced in HD patient and mouse brain and muscle (Chaturvedi et al., 2009, Cui et al., 2006, Chaturvedi et al., 2010), its knockout in mice leads to selective striatal lesions (Lin et al., 2004), and HD striatal neurons expressing exogenous *PGC-1α* are resistant to 3-nitropropionic acid (3-NP) (Weydt et al., 2006).

4.6.2.4 DNA repair

DNA repair pathways were downregulated in HD blood and correlated with disease severity (TMS). GO:6281 includes terms for all major DNA repair pathways, and may represent the downregulation of protective factors such as FAN1, which is included as part of the ICL repair group (Consortium, 2016). These pathways are likely to be relevant to somatic expansion that may influence disease onset and progression (Jonson et al., 2013b, Massey and Jones, 2018, Holmans et al., 2017).

4.6.2.5 Disease severity

The signature of pathway dysregulation identified in HD whole blood correlates with TMS in HD subjects from Track-HD. It also significantly overlaps with that recently found in unstimulated HD monocytes (Miller et al., 2016). This enrichment was driven primarily by upregulation of immune pathways, as might be expected given that Miller et al. (2016) isolated myeloid cells.

4.6.2.6 Comparison with HD brain

To overcome the annotation gap commonly observed with publicly-derived pathway databases and to investigate whether gene expression changes from HD brain are also present in blood, GSEA was performed using brain co-

expression networks derived from HD (Neueder and Bates, 2014) and control (Gibbs et al., 2010, Braineac, 2016) subjects. Several HD brain modules were significantly dysregulated in HD blood, suggesting a common signature of transcriptional dysregulation between blood and brain.

Brain modules upregulated in blood were enriched for immune-related genes, confirming the results of the pathway analysis. Strikingly, two of the modules most significantly dysregulated in HD caudate, 48 (CNpos2) and 66 (CNneg1), were also significantly dysregulated in the same direction in both independent blood datasets. Compared with other brain regions, the caudate has the largest number of expression changes and the highest correlation with HD (Neueder and Bates, 2014). Module 48 (CNpos2), the second most significantly upregulated module in caudate, is enriched for transcriptional regulators, chromatin modifiers and genes involved in mRNA processing (Neueder and Bates, 2014). It is also significantly enriched for immune response genes, giving further support to the pathway results.

Module 66 (CNneg1), the most significantly downregulated module in caudate, contains genes involved in neuronal function, particularly synaptic function and plasticity, and ion channels. Around half of its hub genes are implicated in synaptic function and all were significantly downregulated in Hodges et al. (2006). Though synapses are not present in blood, synaptic genes may be dysregulated in circulating cells without significant pathogenic impact, or alternatively they may serve distinct functions in blood cells. Indeed, Cai et al. (2010) found that the synaptic module was well preserved between brain and blood.

In addition, gene expression and pathway dysregulation from HD prefrontal cortex (Labadorf et al., 2015) was replicated in HD blood. The high degree of overlap increases confidence in the shared signal between blood and brain. A significant proportion of the modules dysregulated in HD blood correlated with TMS.

Mina et al. (2016) performed WGCNA on the Leiden blood sample, finding modules related to immune response that were associated with TFC and TMS. Furthermore, by comparing biological annotations of their HD blood modules with those they derived from Hodges et al. (2006) brain expression data, they showed a common signature between blood and caudate related to immune response. These analyses, using different methodology to those presented here, lend further support to the results above.

The demonstration of a transcriptional signature common to both HD blood and brain supports the use of blood cells to study aspects of HD biology. HD model systems, such as mice, only recapitulate aspects of disease and must be compared to the relevant data in human tissue (Morton and Howland, 2013, Ehrnhoefer et al., 2009). Access to brain tissue is very limited and tends to be from post-mortem subjects with advanced disease, which affects RNA integrity (Montanini et al., 2013, Tomita et al., 2004). Blood, by contrast, is readily available and can be obtained longitudinally from HD subjects.

4.6.3 Comparison with Alzheimer's disease brain

In AD, amyloid plaques are surrounded by chronically activated microglia (Gomez-Nicola et al., 2013, Olmos-Alonso et al., 2016) and GWA studies have identified immune-related genes as risk factors for LOAD (Wyss-Coray and Rogers, 2012). Hong et al. (2016a) showed that early in the disease process, before plaque formation, microglia and complement activation drive synaptic loss, a process that may reflect reactivation of developmental synaptic pruning (Hong et al., 2016b).

In this chapter, in HD blood there was significant upregulation of all four immune modules associated with AD brain in the IGAP GWAS (International Genomics of Alzheimer's Disease, 2015), as well as the most significant immune and microglia-related modules from the Zhang et al. (2013) study of AD brain. In a co-expression network generated from prefrontal cortex of 194 HD patients, Zhang et al. (2013) found that their most significant immune and microglia module was not significantly dysregulated in HD prefrontal cortex and did not correlate with CAG repeat length. This may be because cortex shows less severe pathology and transcriptional dysregulation than caudate (Hodges, 2006).

4.7 Summary

This chapter investigated transcriptional dysregulation in peripheral HD blood of two independent cohorts from Track-HD (Tabrizi et al., 2009b) and Leiden, and compared it to datasets generated from brain in HD and Alzheimer's. There was significant dysregulation of brain Weighted Gene Correlation Network Analysis (WGCNA) modules in the same direction in blood, as well as significant dysregulation of pathways. The transcriptional signature replicated dysregulation seen in HD brain, particularly the caudate which is the tissue most vulnerable to the disease. Immune gene sets were notably upregulated in all analyses and this signal overlapped with the transcriptional signature of Alzheimer's disease (AD) brain. Overlapping immune upregulation in HD and AD suggests these two distinct neurodegenerative diseases share some common pathogenic mechanisms, including macrophage function (Hong et al., 2016a). The strong immune signal is consistent with transcriptional studies in numerous neurodegenerative diseases, indicating a key role for inflammation in neuronal degeneration.

4.8 Publications relating to this chapter

The work presented in this chapter was published in:

Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. Hensman Moss, Davina J.*, **Flower, Michael D.***, Lo, Kitty K., Miller, James R. C., van Ommen, Gert-Jan B., 't Hoen, Peter A. C., Stone, Timothy C., Guinee, Amelia, Langbehn, Douglas R., Jones, Lesley, Plagnol, Vincent, van Roon-Mom, Willeke M. C., Holmans, Peter[#] and Tabrizi, Sarah J.[#] ***Scientific Reports***, 2017 Mar 21;7:44849. doi: 10.1038/srep44849.

* These authors should be regarded as joint first authors.

[#] These authors jointly supervised the work.

Chapter 5 Cell models of *HTT* CAG repeat instability

5.1 Background

5.1.1 Repeat instability

The pathogenic CAG repeat in *HTT* is inherently unstable, and tends to expand over time, particularly in the striatum (Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009). Expansion produces an increasingly toxic polyglutamine protein, and is correlated with earlier onset and increasingly severe disease (Lee et al., 2012d, Telenius et al., 1994, Gomes-Pereira et al., 2001, Fortune et al., 2000, Kennedy and Shelbourne, 2000, Swami et al., 2009), suggesting it is a mechanism underlying the tissue-specific, progressive nature of the disease (Goula et al., 2012, Wheeler et al., 1999, Shelbourne et al., 2007a). Several modifiers of repeat stability have been identified (Wheeler et al., 2007) including environmental stress (Chatterjee et al., 2015), chemical inducers (Gomes-Pereira and Monckton, 2004a) and DNA repair genes, particularly the mismatch repair pathway (Kovalenko et al., 2012). Knockout of *Msh2* or *Msh3* significantly reduces somatic expansion in HD (Manley et al., 1999, Kovtun and McMurray, 2001, Wheeler et al., 2003, Owen et al., 2005), DM1 (van den Broek et al., 2002, Savouret et al., 2003) and fragile X (Lokanga et al., 2014, Zhao et al., 2015b, Zhao et al., 2016) mouse models, suggesting repeat expansion is a result of DNA repair activity, particularly MutSβ (MSH2/MSH3). In HD (Kennedy et al., 2003) and DM1 (Ashizawa et al., 1993) patients and transgenic mouse models (Lia et al., 1998, Fortune et al., 2000, Mangiarini et al., 1997, Kennedy and Shelbourne, 2000), there is no significant link between somatic expansion rate and the proliferative capacity of the tissue, implying expansion occurs during transcription or repair, rather than DNA replication. Integrating somatic mosaicism and transcriptomic data in HD transgenic mice showed a negative correlation between cell cycle pathways and tissue-specific instability, consistent with a cell-cycle independent mechanism (Lee et al., 2010). Whilst DNA replication is limited to S-phase, DNA repair occurs at all stages of the cell cycle and is a strong candidate as the driver of expansion.

The mechanisms that give rise to repeat instability in patients are not yet fully understood, but significant evidence implicates the formation of abnormal DNA secondary structures and DNA damage induced by oxidative stress. Slipped-strand DNA structures may occur when DNA is unwound during transcription or repair (Lopez Castel et al., 2011), though the involvement of DNA replication in proliferative tissues is also possible. In HD, DM1 and SCA7 (Freudenreich et al., 1997, Kang et al., 1995, Liu et al., 2010a, Panigrahi et al., 2002, Cleary et al., 2010, Nenguke et al., 2003), contractions occur when the CTG repeat is on the lagging strand template and expansions when it is CAG, suggesting CAG repeats have a higher propensity to form DNA slip outs or are processed differently by DNA repair machinery. Slipped DNA structures are likely more prone to forming on the lagging strand template because it remains single stranded for relatively long ~300 nt stretches prior to Okazaki fragment synthesis (Hay and DePamphilis, 1982, Anderson and DePamphilis, 1979). Slip outs have been found at CTG repeat tracts in non-mitotic DM1 patient tissue that show somatic expansion, including brain, muscle and heart (Axford et al., 2013), suggesting they are stable, and not merely transient mutagenic intermediates.

HD patient striatum shows more oxidative lesions (Browne et al., 1997), and mouse models accumulate oxidative damage in tissues affected by CAG repeat expansion, such as the liver and brain (Kovtun et al., 2007), specifically at CAG repeat DNA, in a length dependent manner (Bogdanov et al., 2001, Goula et al., 2009). Knockout of the BER glycosylase OGG1,

which removes 8-oxoG, reduces somatic expansion in R6/1 mice (Kovtun et al., 2007). In Friedreich's ataxia, where GAA repeats are also unstable and the absence of MMR components *Msh2* or *Msh6* accelerates expansion in mouse models (Bourn et al., 2012, Lai et al., 2014, Krasilnikova and Mirkin, 2004), frataxin deficiency is directly associated with increased cellular oxidative stress in patients (Calabrese et al., 2005, Armstrong et al., 2010).

The existence of rare DM1, HD and SMBA families with consistent contractions rather than expansions, the bias towards contraction in CAG and CTG expansion models where MMR factors such as *MSH2* and *MSH3* have been inactivated (Dragileva et al., 2009, Foiry et al., 2006, Kovtun et al., 2004, Manley et al., 1999, Savouret et al., 2003, Savouret et al., 2004, van den Broek et al., 2002, Wheeler et al., 2003), and the prevalence of CTG contractions through the female germline of *LigI* deficient DM1 mice all suggest the mechanisms underlying expansion and contraction may be distinct (Slean et al., 2016).

5.1.2 Cell models of repeat instability

5.1.2.1 Huntington's disease

Few cell models of robust and significant *HTT* CAG repeat instability in the absence of genotoxic stress are available. Kovtun et al. (2007) exposed patient-derived fibroblasts (FB) and lymphoblasts (LB) with 69 CAG repeats to H_2O_2 up to three times over nine days and showed an increase of 1 CAG. Cannella et al. (2009) cultured 58 HD patient-derived LB lines with repeat lengths between 39 and 120 CAG for 6-12 months. Those with over 64 CAG expanded by an average 3 repeats over 6 months (53.22 ± 16.29 days/Q). They found that treatment with the DNA intercalator ethidium bromide, the GC/AT modifier ethylmethanesulphonate (EMS), or the interstrand crosslinker mitomycin C (MMC), narrowed the traces and induced some negative skew in two lines with 74 and 80 CAG repeats.

Jonson et al. (2013a) generated pluripotent embryonic stem cells with 127 CAG repeats from R6/1 HD mice and showed that exposure to chronic oxidative stress with H_2O_2, which induces single and double strand breaks, accelerated CAG repeat expansion up to threefold. In undifferentiated cells, expansion rate increased from 2 to 5 CAG repeats over 12 passages (5 weeks), and in differentiating cells it increased from 0.8 to 1.5 repeats over 12 days. Potassium bromide ($KBrO_3$), which generates oxidative DNA damage, particularly 8-oxoG (Ballmaier and Epe, 2006), induced a more modest acceleration. Methyl methanesulfonate (MMS), an alkylating agent that methylates DNA (Sanderson and Shield, 1996), had no effect, suggesting that oxidative rather than alkylating damage promotes expansion.

Jacquet et al. (2015) showed that human HD embryonic stem cells with 38-43 CAG repeats are stable in culture and during cardiomyocyte differentiation, whereas Mollica et al. (2016) reported that 44Q patient-derived fibroblasts showed a small positive skew in the electrophoresis trace over 35 days, which was reduced by the DNA methyltransferase inhibitor 5-azacytidine.

5.1.2.2 Myotonic dystrophy

5.1.2.2.1 Lymphoblastoid cells

Repeat instability appears to occur more readily in myotonic dystrophy type 1 (DM1) cells, with patient-derived lymphoblasts (LB) having long been known to show expansion in culture (Ashizawa et al., 1996, Bidichandani et al., 1999, Ashizawa et al., 1993).

5.1.2.2.2 Artificial cell models

Human fetal lung fibroblasts stably transfected with a CTG repeat-containing plasmid which was not expressed due to a 5' transcription terminator showed expansion that was accelerated when transcription was activated by Cre-mediated excision of the terminator, suggesting transcription could contribute to instability (Nakamori et al., 2011). Experiments in HEK 293 cells transfected with CTG-containing plasmids suggested 5' oxidative damage increased instability and 3' lesions led to contraction (Lai et al., 2013). In human fibrosarcoma cells transfected with an 800 CTG plasmid, knockdown of *MSH2* and *MSH3* reduced expansion (Nakatani et al., 2015a). In a human foetal astrocytic line, knockdown of *MSH3* or inactivation of its ATPase activity reduced expansion of a non-pathogenic CTG repeat, whereas *MSH3* overexpression increased it (Keogh et al., 2017).

5.1.2.2.3 Stem cells

Human embryonic stem cells (hESC) derived from DM1 patients with 370 or 1800 repeats showed instability that was stabilised by outer plexiform layer (OPL) differentiation and downregulation of MMR genes, but HD lines were stable (Seriola et al., 2011a, De Temmerman et al., 2008). In another study, DM1 iPSCs showed instability at a rate that correlated with repeat length and was reduced by shRNA-mediated *MSH2* knockdown or differentiation into embryoid body or neurospheres, whereas once again HD iPSCs were stable (Du et al., 2013a).

5.1.2.2.4 Animal cells

In cells derived from a transgenic DM1 mouse, there was no correlation between expansion and mitotic rate, suggesting DNA replication is insufficient to drive expansion (Gomes-Pereira et al., 2014a, Gomes-Pereira and Monckton, 2004b, Gomes-Pereira et al., 2001). When the cell cycle was chemically or genetically arrested, the repeat continued to expand at the same rate, supporting a cell division-independent mutational pathway.

5.1.2.2.5 Friedreich's ataxia

HEK 293 cells stably transfected with a GAA construct showed expansion in culture, which was decreased by reducing transcription (Ditch et al., 2009) or knockdown of *MSH2* or *MSH3* (Halabi et al., 2012a). In patient-derived LBs, the BER-inducing alkylating agent temozolomide led to GAA contraction (Lai et al., 2014). In fibroblasts, ectopic expression of *MSH2* and *MSH3* led to expansion and shRNA-mediated knockdown of either reduced it (Halabi et al., 2012b). FRDA iPSCs showed instability (Du et al., 2012a, Ku et al., 2010) which was reduced by *MSH2 (Du et al., 2012b, Ku et al., 2010)*, *MSH3 (Ku et al., 2010)* or *MSH6 (Du et al., 2012b, Ku et al., 2010)* knockdown and differentiation into neural stem cells (Du et al., 2012b, Ku et al., 2010).

5.1.2.2.6 SCA10

LB cells derived from SCA10 patients show instability (Lin and Ashizawa, 2003), and when HeLa cells were complemented with an ATTCT repeat they demonstrated length dependent expansion (Liu et al., 2007).

5.1.3 Stem cells

HD patient-derived stem cells are a valuable system in which to study pathogenesis because they can be expanded indefinitely, retain the potential to differentiate into neurons, and can generate specified cell populations, including DARPP-32 positive striatal MSNs, thereby providing a physiologically-relevant model of HD (Aubry et al., 2008). They show a robust phenotype with expression, electrophysiological, metabolic and cellular adhesion changes that correlate

with pathogenic *HTT* CAG repeat length and recapitulate aspects of cellular phenotypes found in HD patients and mice (Consortium, 2012, Consortium, 2017). Human-derived stem cells are also currently the focus of cell replacement therapies for HD, having shown improvement in neurogenesis, immune dysfunction, mitochondrial function and cell survival in animal models (Rosser and Bachoud-Levi, 2012, Maucksch et al., 2013, Connor, 2018).

5.1.3.1 Medium spiny neurons

GABAergic medium-sized spiny neurons (MSN) are the principal projection neurons of the striatum which specifically degenerate early in HD (Lange et al., 1976, Reiner et al., 1988, Gerfen, 1992, Ouimet et al., 1984), and therefore represent the most physiologically relevant cell model in which to study repeat expansion. *In vivo*, they derive from the lateral ganglionic eminence (LGE), likely driven by activin A, a TGFβ family protein that induces forebrain neurogenesis (Arber et al., 2015).

The Arber et al. (2015) MSN differentiation protocol induces expression of striatal MSN markers including CTIP2, DARPP-32, NOLZ1, GSX2, FOXP2, DLX2, ARPP21, CALB1, PENK, TAC1, GAD1, DRD1 and DRD2. Straccia et al. (2015) assessed human foetal WGE (whole ganglionic eminence, the striatal priodrium) and cortex, and adult caudate, putamen and motor cortex for expression of a panel of genes involved in striatal development. **WGE** was distinguished by high expression of DLX1, DLX5, DLX6, EBF1, LHX6, NKX2-1, and SIX3. In **mature striatal neurons** there was specific upregulation of ADORA2A, CALB1, DRD1, DRD2, PENK, and TAC1, with comparably increased expression of GAD2 and OPRM1 also occurring in foetal WGE, and downregulation of WGE markers DLX1, DLX5, DLX6, EBF1, LHX6, and SIX3. They concluded that **direct** (TAC1, DRD1) and **indirect** (PENK, DRD2) striatal pathway genes were the most specific markers for adult MSNs, together with CALB1 and ADORA2A. DARPP-32 (PPP1R1B) expression is often used to identify MSNs, but interestingly whilst it was increased, it was not a reliable marker of mature striatal neurons. CTIP2 (BCL11B), another traditional MSN marker, actually decreased relative to foetal WGE, which parallels data in the Human Brain Atlas (Sunkin et al., 2013). Therefore, CTIP2 and DARPP-32 expressing neurons may represent the foetal stage of striatal development, which would explain their classification as an MSN markers in studies where foetal brain tissue or differentiated stem cells were analysed (Delli Carri et al., 2013).

5.2 Aims

Somatic instability cannot be monitored longitudinally in human brain cells, but age-dependent, expansion-biased, tissue-specific somatic mosaicism has been replicated in transgenic mouse models of HD (Kennedy and Shelbourne, 2000, Mangiarini et al., 1996, Mollersen et al., 2010, Gonitel et al., 2008, Kovtun and McMurray, 2001) and DM1 (Seznec et al., 2000). Animal studies of somatic instability are limited by the complex nature of tissues, which are made up of multiple cell types with differing proliferative capacities, and the inability to determine the replicative history of any given cell *in vivo* (Gomes-Pereira et al., 2014a). Biochemical studies have largely used cell-free extracts or purified DNA repair proteins to study activity at trinucleotide repeats, though it is not possible to recapitulate the complex interactions of the DNA repair network in cell free systems (Stevens et al., 2013). Cell models have the advantage of lower complexity than animal models and facilitate the modulation of DNA repair proteins (Nakatani et al., 2015a). Current HD cell models lack significant expansion, requiring long periods in culture or chronic genotoxic insult to produce subtle repeat length change, which has limited our understanding of the molecular basis of repeat instability (Gomes-Pereira et al., 2014a).

To explore the role of DNA repair in mediating instability, this chapter sets out to a generate cell culture models that reliably reproduce the time-dependent and expansion biased repeat instability seen in HD patients and mouse models. Once established, it investigates the effect of genotoxic stress and the role of DNA repair proteins in repeat stability in different cell types and differentiation states.

5.3 Methods

5.3.1 Cell viability assays

Cell death was measured either by LDH cytotoxicity kit (Promega cat #G1780) or MTT assay 24 h following stress.

5.3.1.1 Cytotoxicity kit

Cells were diluted to the appropriate concentration, added to a 96 well plate and exposed to genotoxic stress at the indicated concentration in a 100 µL volume per well. Cells were washed, 100µL fresh media added, and then incubated at 37°C for 24 h. 10 µL of 10X lysis solution was added to positive control wells and incubated for 45 min at 37°C. 50 µL aliquots of media from all wells were transferred to a clear bottom plate and 50 µL of CytoTox reagent added to each. The plate was covered to protect from light and incubated at room temperature for 30 min. 50 µL of Stop solution was added to each well and the absorbance was measured at 490 nm on a Tecan sunrise absorbance microplate reader. The average values of the negative control (medium only) wells was subtracted from all wells, then percentage cytotoxicity was calculated relative to positive control wells.

5.3.1.2 MTT assay

The MTT assay is a colorimetric readout of cellular metabolic activity used to measure cytotoxicity by loss of viable cells. Cellular oxidoreductase enzymes reduce the tetrazolium dye 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) to insoluble formazan, which is purple. Cells were cultured in a 96 well plate at 10,000 cells per well in 100 µL of medium overnight to adhere. The selected genotoxin was added at the appropriate concentration and duration. Media was changed and MTT added at 5 mg/ml in PBS. Cells were incubated at 37°C for 1 h, then media was removed and formazan crystals were dissolved by the addition of 50 µL of DMSO and incubation at room temperature for 30 min. The plate was read on a microplate absorbance reader at 570 nm.

5.3.2 Oxidative stress

Hydrogen peroxide (H_2O_2), which oxidises bases and induces single and double strand breaks (Driessens et al., 2009), was added to media at the required concentration by serial dilution from a 1000 µM stock and incubated for 30 min (Jonson et al., 2013a), then washed once with PBS and normal culture media added.

5.3.3 shRNA-mediated FAN1 knockdown

An shRNA hairpin targeting endogenous *FAN1* (target sequence GTAAGGCTCTTTCAACGTA), synthesised by GeneArt, was subcloned into the pSUPER.retro.puro vector (see Appendix) and transfected into Phoenix Ampho packaging cells using Lipofectamine LTX (Gandhi et al., 2009, Wood-Kaczmar et al., 2008). After 16h, 8 ml fresh media was added, then media containing mature retrovirus was harvested 48h post transfection. This was filtered and frozen at -80°C or used directly. It enabled the long term, stable knockdown needed to test FAN1 involvement in CAG repeat instability in culture and during neuronal differentiation.

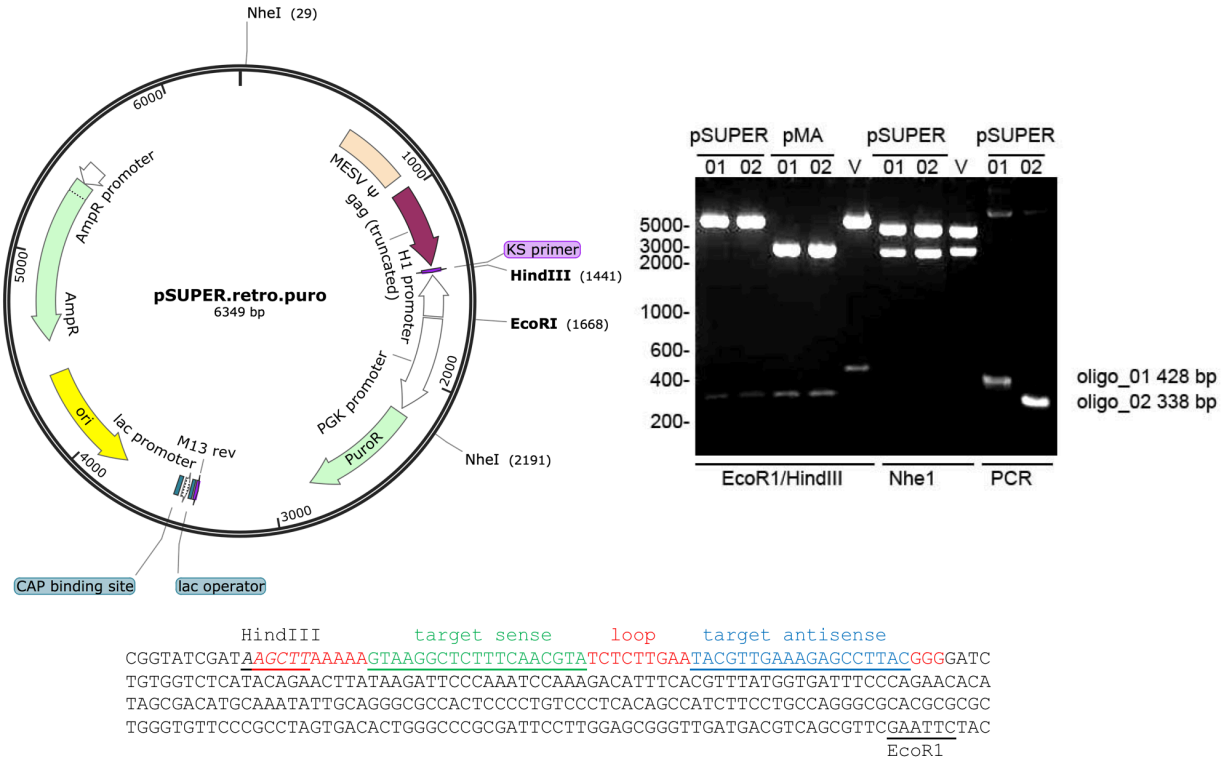


Figure 5.1. Cloning FAN1 shRNA into the pSUPER.retro.puro vector.

Top left *– pSUPER.retro.puro vector map showing HindIII, EcoR1 and NheI restriction sites.* ***Top right*** *– ligation of shRNA oligonucleotide into pSUPER.retro.puro. The shRNA used in this study is indicated by '02'. The insert of around 280 bp was excised by HindIII and EcoR1 cuts from the GeneArt synthesised plasmid (pMA) and pSUPER.retro.puro clones. Ligation was confirmed by PCR using forward primer ACGGCACCTTTAACCGAGAC within pSUPER and reverse primer AGGCTCTTTCAACGTATCTCTTGA within the shRNA oligo, giving a 338 bp product.* ***Bottom*** *– GeneArt fragment sequence showing shRNA region, and HindIII and EcoR1 sites.*

Media containing retrovirus that encodes shRNA targeting FAN1 or empty vector was mixed one to one with normal iPSC media and supplemented with polybrene (8 µg/ml). This media was added to iPSCs at approximately 70% confluence and incubated for 16 h. Cells were washed with PBS and fresh media added for 48h prior to selection with puromycin (1 µg/ml). Media was changed on at least alternate days, monitoring to minimise the number of dead cells in the culture. Colonies of transduced cells were detected after 10-14 days. Untreated cells were cultured alongside the selected cells and used as controls in subsequent experiments. A similar protocol was trialled three times with 125Q lymphoblastoid (LB) cells, but none survived selection, indicating failure of transduction.

5.3.4 Immunofluoresence

Antibodies used in this study are given in the table below.

Antibody	Target	Manufacturer	Cat #	Type	Species	Dilution
Primary	DARPP-32	Santa Cruz	sc-11365 (H-62)	Polyclonal IgG	Rabbit	1:200
	CTIP2	Abcam	ab18465 (25B6)	Monoclonal IgG2a	Rat	1:200
	βIII tubulin	Abcam	ab107216	Polyclonal IgG	Chicken	1:500
Secondary	Goat anti-Rabbit Alex Fluor 488	Invitrogen	A-11008	Polyclonal IgG	Goat	1:1000
	Goat anti-Rat Alexa Fluor 568	Invitrogen	A-11077	Polyclonal IgG	Goat	1:1000
	Goat anti-Mouse Alex Fluor 647	Invitrogen	A-21236	Polyclonal IgG	Goat	1:1000

Table 5.1. Antibodies for immunofluorescence.

5.3.5 Quantative real time PCR (qPCR)

The following Taqman probes (Thermo) were used to assess FAN1 knockdown and MSN differentiation in iPSC lines. Housekeeping genes used in the 2^-ΔΔCt analysis were *ACTB*, *ATP5B*, *EIF4A2* and *SDHA*.

Gene	Probe
ACTB	Hs01060665_g1
ADORA2A	Hs00169123_m1
ATP5B	Hs00969569_m1
BCL11B (CTIP2)	Hs01102259_m1
CALB1	Hs01077197_m1
DRD1	Hs00265245_s1
DRD2	Hs00241436_m1
EIF4A2	Hs00756996_g1
FAN1	Hs00429686_m1
GAD2	Hs00609534_m1
HTT	Hs00918174_m1
OPRM1	Hs01053957_m1
PENK	Hs00175049_m1
PPP1R1B (DARPP-32)	Hs00259967_m1
PPP1R1B (DARPP32)	Hs00259967_m1
SDHA	Hs00188166_m1
TAC1	Hs00243225_m1

Table 5.2. Taqman qPCR probes.

5.4 Contributions

ReNeuron neural stem cells (NSC) were transduced with the 129Q vector as described in Trager et al. (2014), and baseline characterisation of cellular phenotype was conducted by Rhia Ghosh (UCL) and Alison Wood-Kaczmar (UCL). shRNA vector cloning and western blotting was conducted by Rob Goold (UCL). Blood sampling and peripheral blood mononuclear cells (PBMC) preparation were performed by Michael Flower. 125Q lymphoblastoid (LB) cell transformation was conducted by European Collection of Authenticated Cell Cultures (ECACC) lab at Public Health England (PHE), Salisbury, and iPSCs were generated by Censo Biotechnologies, Midlothian. All cell culture, cloning, retroviral transduction, genotoxic stress, CAG repeat sizing, immunofluorescence, microscopy and data analysis was conducted by Michael Flower. These results have been published in Goold et al. (2018).

5.5 Results

5.5.1 ReNcell neural stem cells

5.5.1.1 ReNcell VM 129Q

5.5.1.1.1 Hydrogen peroxide titration

ReNcell VM neural stem cells were lentivirally transduced with *HTT* exon 1 containing either non-pathogenic 29, or pathogenic 71 or 129 CAG repeats. Each line was neuronally differentiated as in Donato et al. (2007), with mitotic neural stem cells cultured in parallel. Those expressing 129 repeats were found to be more sensitive to oxidative stress during, rather than after differentiation with 25% cell death induced by approximately 25 µM or 250 µM hydrogen peroxide respectively.

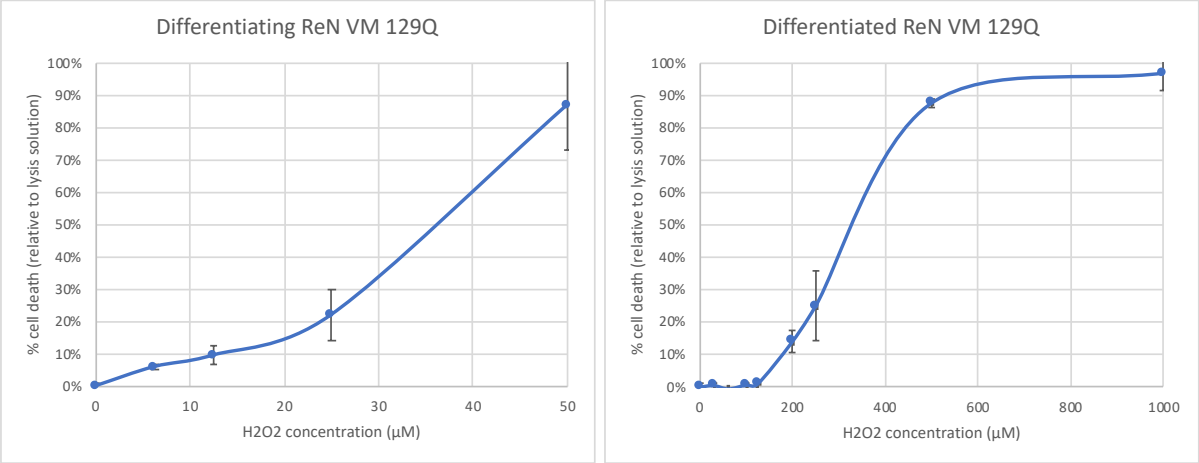


Figure 5.2. Oxidative stress in ReNeuron VM 129Q cells.
Left *– cells 3d into the differentiation protocol stressed with the indicated H_2O_2 concentration for 30 min.* ***Right*** *– differentiated cells 17 days from neuronal induction. 3 independent replicates for each data point, error bars represent SEM. Cell death measured by LDH cytotoxicity assay 24h after stress.*

5.5.1.1.2 CAG repeat sizing

Differentiation experiments were performed at least in triplicate and samples taken serially for CAG repeat sizing. At baseline, cells transduced with the 29Q vector were sized at 30.8 ± 0.0028Q, the 71Q vector at 70.3 ± 0.13Q and the 129Q vector at 132.1 ± 1.46Q. There was no significant change in modal CAG repeat length, proportional expansion analysis or somatic instability index, either in culture as NSCs or during neuronal differentiation over 8 weeks. Representative fragment analysis traces are given below.

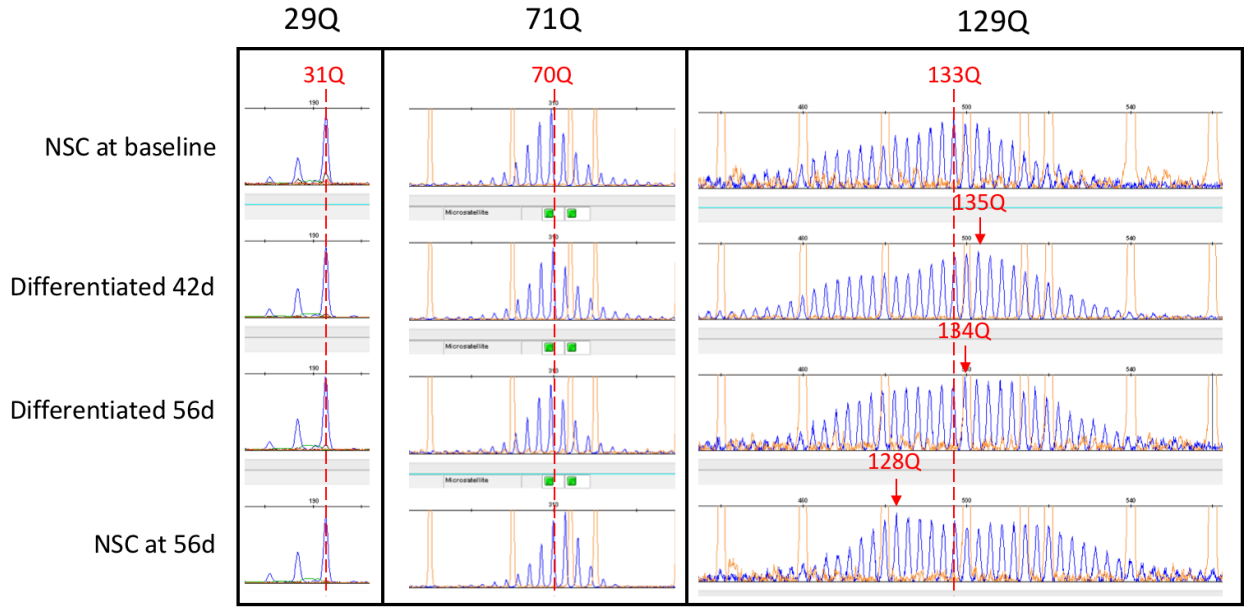


Figure 5.3. Representative CAG repeat sizing from ReN VM neural stem cells (NSC) differentiated for 56 days. *__Top__ – NSC at baseline. __Middle__ – neuronally differentiated for 42 and 56 days. __Bottom__ – NSC cultured in parallel for 56 days. __Left__ – ReN VM transduced with non-pathogenic 29Q HTT exon 1. __Middle__ – 71Q HTT exon 1. __Right__ – 129Q HTT exon 1. Modal CAG repeat length at baseline is given as a red dotted line. Modal CAG length at other times represented by red arrow. Q – modal CAG length.*

HTT exon 1 repeat length	Treatment	Modal CAG length	Change in modal CAG	Proportional expansion	Instability index
71Q	NSC baseline	70.1 (± 0)	0 (± 0)	0.485 (± 0.004)	0 (± 0)
	NSC d56	70.75 (± 0.37)	0.65 (± 0.367)	0.675 (± 0.012)	0.785 (± 0.02)
	Differentiated d56	70.15 (± 0.04)	0.05 (± 0.041)	0.505 (± 0.102)	0.32 (± 0.261)
129Q	NSC baseline	130.53 (± 1.35)	0 (± 0)	0.557 (± 0.09)	0 (± 0)
	NSC d56	128.63 (± 0.68)	-1.9 (± 1.429)	0.590 (± 0.072)	0.867 (± 1.38)
	Differentiated d56	130.5 (± 1.5)	-0.033 (± 0.984)	0.613 (± 0.07)	1.11 (± 0.582)

Figure 5.4. Repeat expansion analysis in ReN VM cells cultured as neural stem cells (NSC) or differentiated (MSN) for 56 days. *Change in modal CAG is given relative to baseline. For proportional expansion, 0.5 represents a normal distribution with a mode equal to the baseline mode. The maximum is 1.0 (the entire distribution is greater than the control mode) and minimum is 0.0 (the entire distribution is less than the control mode). Instability index is given relative to baseline and is measured in CAG units. NSC – neural stem cell. Differentiated – neuronal differentiation. Values are the mean of at least 3 replicates (± SEM).*

NSCs and differentiating or differentiated neurons were then exposed to chronic oxidative stress with H_2O_2. Sublethal doses were used, aiming to kill less than 10% of cells, given that longitudinal sampling from post-mitotic cultures is required (see Methods). Differentiating cells received either 0, 6.25 or 12.5 µM H_2O_2 for 30 min weekly and samples were taken for CAG sizing at 48d form neuronal induction. Differentiated cells were challenged with 0, 75 or 150 µM H_2O_2 for 30 min weekly from day 15 after neuronal induction and samples were taken at 44d from induction. NSCs were cultured and stressed in parallel, and all experiments were conducted in triplicate. There was no significant change in modal CAG repeat length or somatic instability index in any group.

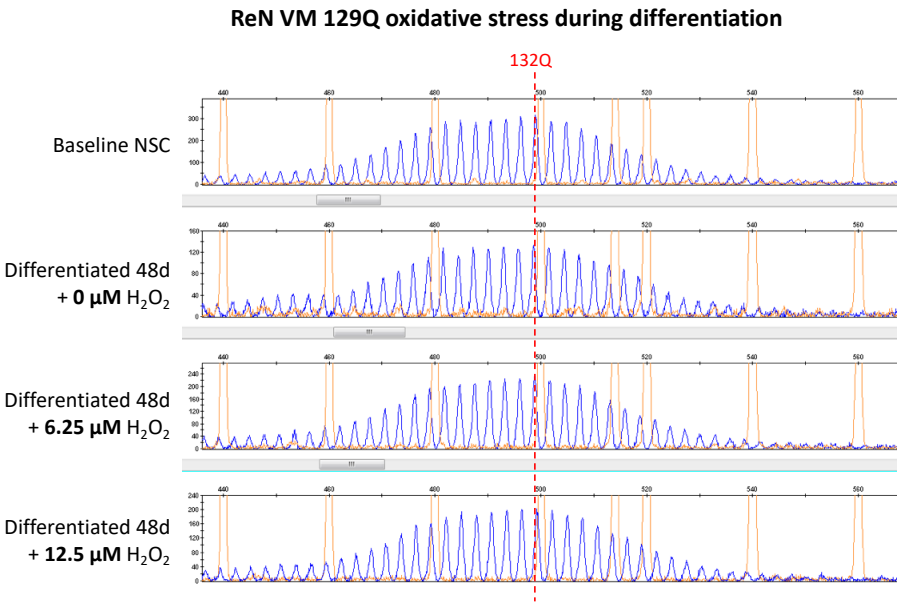


Figure 5.5. Representative CAG repeat sizing from ReN VM 129Q neural stem cells (NSC) chronically stressed with H_2O_2 during differentiation for 48 days.

***Top** – NSC at baseline. **Lower 3 traces** – stressed for 30 min once weekly with the indicated H_2O_2 concentration during neuronal differentiation for 48d. Modal CAG repeat length at baseline is given as a red dotted line. Q – modal CAG length.*

ReN VM 129Q NSC stress during differentiation	Modal CAG length	Change in modal CAG length	Proportional expansion	Instability index
Baseline NSC	132.33 (± 0.55)	0 (± 0.549)	0.358 (± 0.002)	0 (± 0.152)
Differentiated 48d + 0µM H2O2	130 (± 1.88)	-2.33 (± 1.877)	0.335 (± 0.03)	-1.186 (± 0.511)
Differentiated 48d + 75µM H2O2	130.12 (± 1.32)	-2.211 (± 1.322)	0.376 (± 0.019)	0.256 (± 0.136)
Differentiated 48d + 150µM H2O2	129.81 (± 1.7)	-2.521 (± 1.697)	0.389 (± 0.007)	-0.166 (± 0.407)

Table 5.3. Repeat expansion analysis in ReN VM 129Q NSCs chronically stressed with H_2O_2 during differentiation for 48 days.
Change in modal CAG, proportional expansion and instability index are given relative to baseline. NSC – neural stem cell. Differentiated – neuronal differentiation. Values are the mean of at least 3 replicates (± sem).

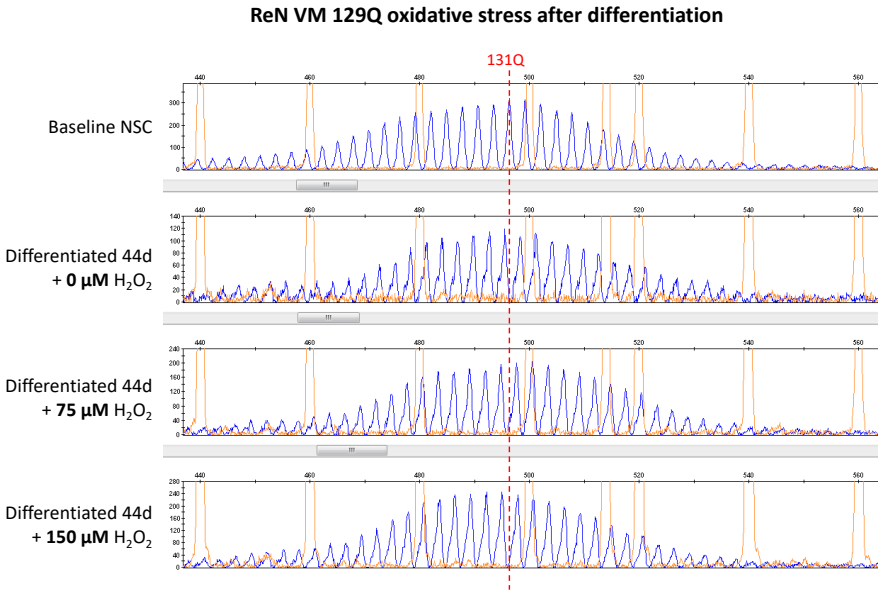


Figure 5.6. Representative CAG repeat sizing from ReN VM 129Q NSCs 44d after initiation of differentiation, chronically stressed with H_2O_2 from day 15.

***Top** – NSC at baseline. **Lower 3 traces** – stressed for 30min once weekly with the indicated H_2O_2 concentration from day 15. Modal CAG repeat length at baseline is given as a red dotted line. Q – modal CAG length.*

ReN VM 129Q NSC stress after differentiation	Modal CAG length	Change in modal CAG length	Proportional expansion	Instability index
Baseline NSC	131.33 (± 0.61)	0 (± 0.606)	0.331 (± 0.021)	0 (± 0.274)
Differentiated 44d + 0µM H_2O_2	130.66 (± 0.51)	-0.672 (± 0.512)	0.415 (± 0.008)	1.461 (± 0.209)
Differentiated 44d + 75µM H_2O_2	132.83	1.50	0.44	1.41
Differentiated 44d + 150µM H_2O_2	131.4 (± 1.33)	0.07 (± 1.33)	0.4 (± 0.031)	0.825 (± 0.493)

Table 5.4. Repeat expansion analysis in ReN VM 129Q NSCs 44d after initiation of differentiation, chronically stressed with H_2O_2 from day 15.

Values are the mean of at least 3 replicates (± SEM), except 75 µM where repeat sizing failed for two replicates.

5.5.1.2 ReNcell CX 129Q

ReN CX NSC electrophoresis traces show a broad, multimodal distribution, with dominant peaks at 83 and 127Q. In the first instance cells were differentiated using the Donato et al. (2007) protocol. Differentiated neurons were chronically stressed with 100 µM H_2O_2 weekly from day 15 and samples taken for repeat sizing at day 49 from neural induction.

Over 7 weeks the smaller peak at 82.8Q increased by 0.96 repeats (proportional expansion +0.15, instability index +1.05Q) and 1.04 repeats (proportional expansion +0.36, instability index +1.34Q) in the absence and presence of chronic oxidative stress respectively.

The larger peak at 126.8Q appeared to increase with differentiation in the absence or presence of oxidative stress by 39.1 repeats (proportional expansion +1.0, instability index +24.43Q) and 37.3 repeats (proportional expansion +1.0, instability index +30.6Q), with none of the original modal allele remaining. The change at the broad 127Q peak is suggestive of expansion, but may represent the selective loss of cells expressing shorter alleles, perhaps because CAG expansion reduces *HTT* expression, thereby providing a selection advantage (Dragatsis et al., 2009). Over the 7 weeks, the smaller 83Q peak showed a increased by 1Q in both mode and instability index.

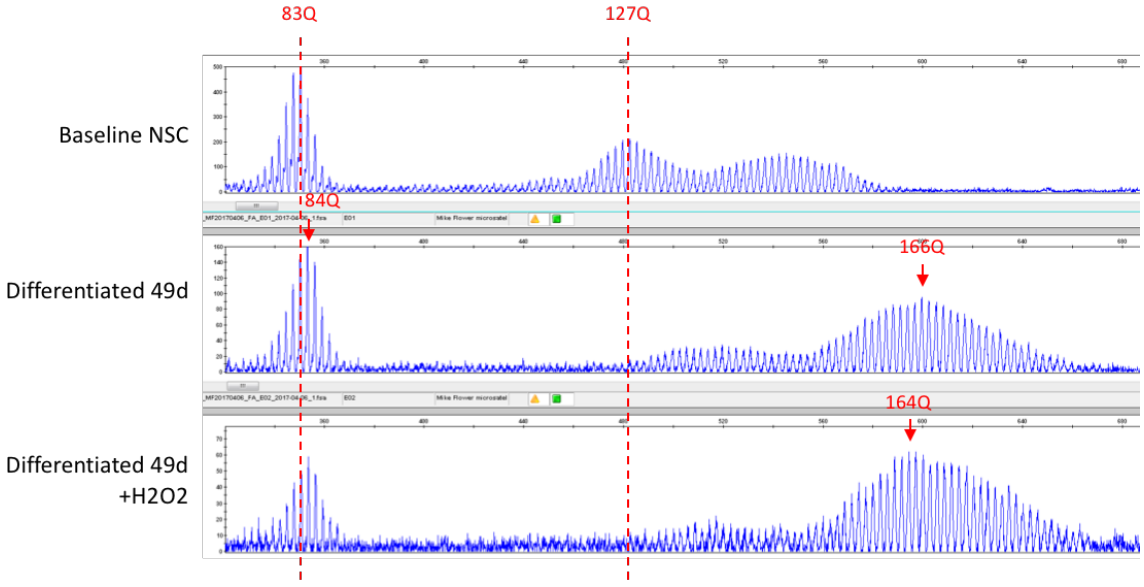


Figure 5.7. Representative CAG repeat sizing in ReN CX 129Q cells differentiated in the presence of chronic oxidative stress.
***Top** – baseline neural stem cells (NSC). **Middle** – neuronally differentiated for 49 days. **Bottom** – neuronally differentiated with 100 µM H_2O_2 stress for 30 min weekly from day 15. Modal CAG repeat length at baseline is given as a red dotted line. Q – modal CAG length.*

However, the ReN CX cells rarely successfully differentiated without losing *HTT* exon 1 expression, demonstrating the toxicity of exon 1 fragment. A later successful differentiation in a sample expressing a peak at 116.3Q baseline is shown

below. At day 5 it had increased by 0.71Q (proportional expansion -0.05, instability index -0.10Q) and at day 14 had increased by 1.00Q (proportional expansion +0.08, instability index +0.91Q).

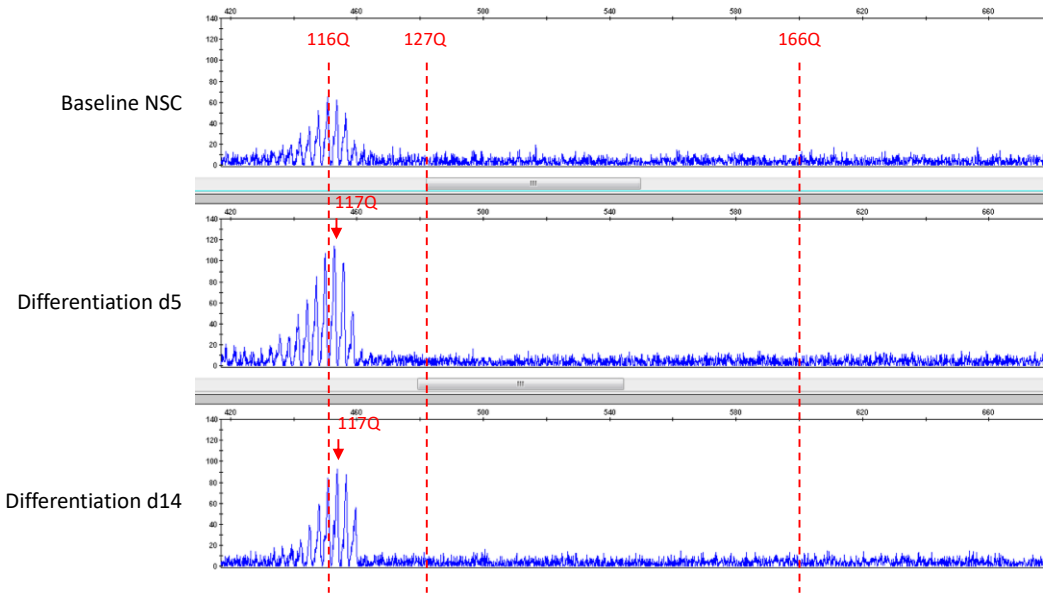


Figure 5.8. Representative CAG repeat sizing in ReN CX 129Q cells differentiated for 14 days.
Top *– baseline neural stem cells (NSC).* ***Middle*** *– neuronally differentiated for 5 days.* ***Bottom*** *– neuronally differentiated for 14 days. Red dotted lines represent the modal CAG length at baseline (116Q), and the modal CAG lengths observed in previous differentiations (127Q and 166Q). Modal CAG length at other times represented by red arrow. Q – modal CAG length.*

5.5.1.3 ReN VM 129Q single cell cloning

To increase sensitivity in the expansion analysis, the polyclonal ReN VM 129Q cells were single cell cloned by serial dilution at 100 or 10 cells/well in 96 well plates (see Chapter 2). Cultures derived from 100 cell/well dilutions contained several distinct peaks, but those from the 10 cell/well dilution produced a single narrow, normally distributed peak. 23 clones were generated from 10 cell/well dilution, two of which were cultured and the CAG repeat sized. Notably both had CAG repeat lengths derived from the top 5% of the original polyclonal distribution (mean 142.3 $\pm$ 1.36Q), again suggesting positive selection for longer repeat length. The mean width of the CAG traces was successfully reduced from 39.34 $\pm$ 1.40Q to 8.09 $\pm$ 4.32Q.

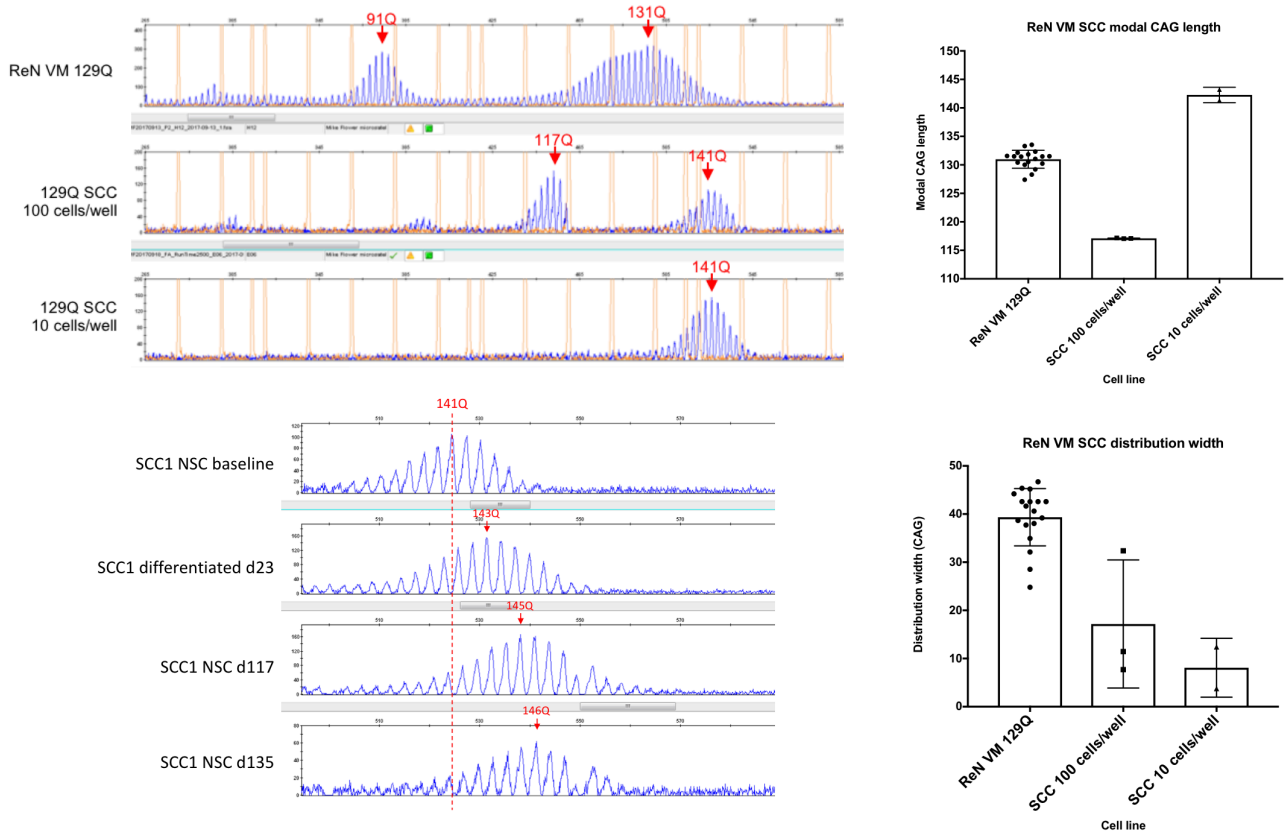


Figure 5.9. CAG repeat sizing in ReN VM 129Q single cell clones (SCC).

Top left *– first is the original polyclonal line, in the middle is cloning with a dilution of 100 cells/well and finally dilution at 10 cells/well.* ***Top right*** *– modal CAG repeat length.* ***Bottom right*** *– width of the CAG repeat distribution on capillary electrophoresis. Note this includes several distinct peaks in the polyclonal and 100 cells/well dilution. Values given are mean ±SEM. ReN VM 129Q – original polyclonal population. SCC– single cell cloning by serial dilution to 100 or 10 cells/well in a 96 well plate format.* ***Bottom left*** *– CAG repeat sizing from a ReN VM 129Q single cell clone in culture and during neuronal differentiation. First is baseline neural stem cells (NSC), second is neuronally differentiated from day 0 and sized at day 23, third is NSC in culture for 117 days, and at the bottom are NSC in culture for 135 days. Modal CAG repeat length at baseline is given as a red dotted line. Modal CAG length on other dates are represented by red arrows. Q – modal CAG length, SCC1– single cell clone number 1.*

Single cell clone	Treatment	Modal CAG length	Change in modal CAG	Proportional expansion	Instability index
SCC1	NSC baseline	140.79	0.00	0.52	0.00
	Differentiated d23	143.11	2.32	0.74	1.82
	NSC d117	145.36	4.57	0.88	4.29
	NSC d135	146.45	5.66	1.00	5.22
SCC2	NSC baseline	141.31	0.00	0.43	0.00
	NSC d14	143.23	1.92	1.00	3.59

Table 5.5. CAG repeat sizing analysis for two single cell clones in culture and during differentiation.
SCC – single cell clones 1 and 2.

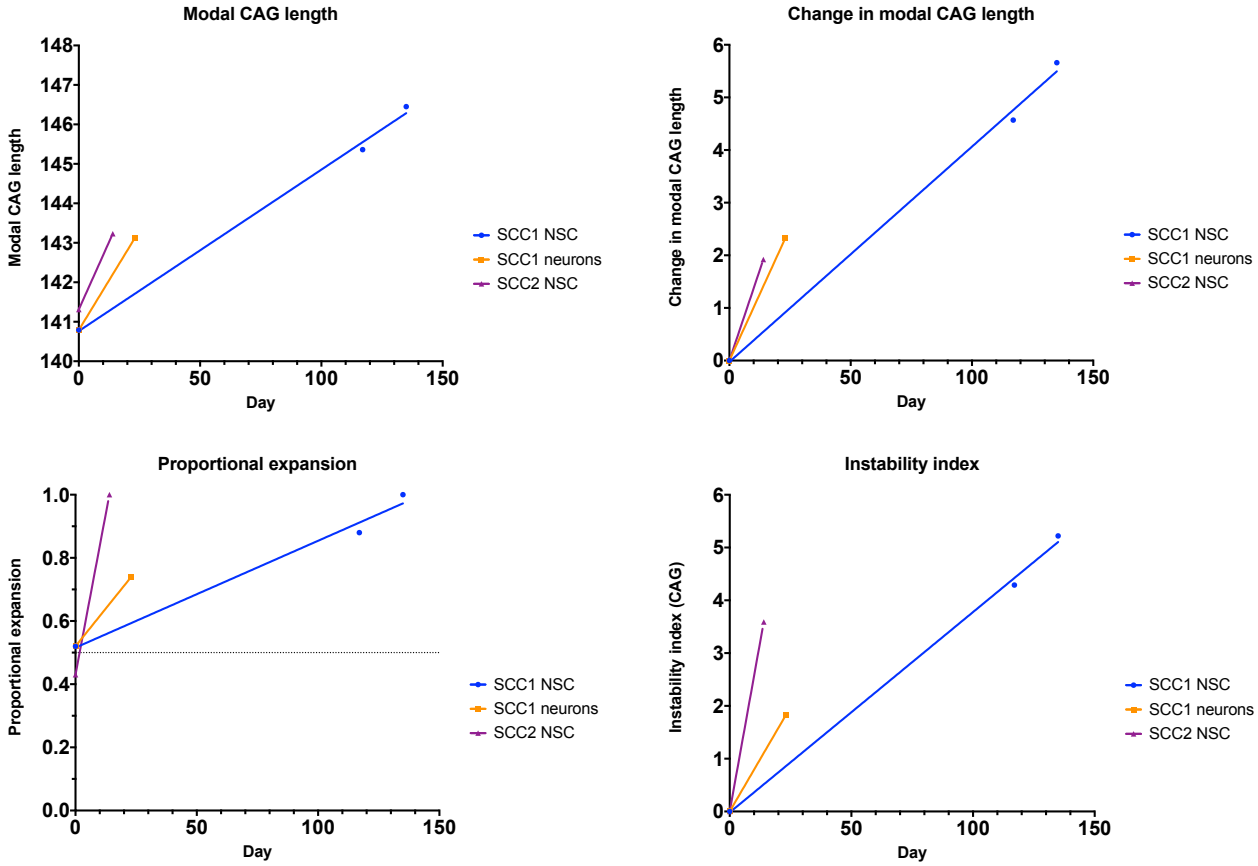


Figure 5.10. CAG expansion analysis of ReN VM 129Q single cell clones.
Top left *– modal CAG repeat length over time.* ***Top right*** *– change in modal CAG repeat length relative to NSC at baseline.* ***Bottom left*** *– proportional expansion analysis relative to baseline.* ***Bottom right*** *– instability index relative to baseline, measured in CAG repeat units. SCC – single cell clones 1 and 2, NSC – neural stem cell. Neurons – neuronal differentiation.*

5.5.2 Track-HD patient-derived lymphoblastoid cells

Cannella et al. (2009) demonstrated *HTT* CAG repeat expansion in patient-derived lymphoblasts (LB) with at least 64 CAG repeats cultured for 6 months. Jonson et al. (2013a) found that chronic oxidative stress with 50 or 150 µM hydrogen peroxide (H_2O_2), which induces an array of DNA damage including oxidised bases and DNA strand breaks (Spencer et al., 1995, Spencer et al., 1996), for 30 min before each passage accelerated expansion of a 127Q *HTT* exon 1 CAG repeat in R6/1 mouse embryonic stem cells in a dose dependent manner (Mangiarini et al., 1996).

5.5.2.1 Hydrogen peroxide titration

HD LBs with 43 CAG repeats, assayed using the MTT method, showed around 25% cell death at 50 µM and 50% at 150 µM H_2O_2.

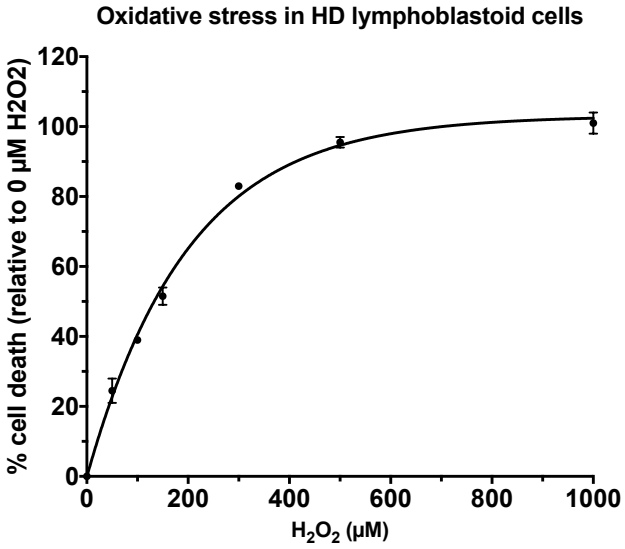


Figure 5.11. Oxidative stress in HD lymphoblastoid (LB) cells.
43 CAG repeat LB cells were exposed to the indicated H_2O_2 concentration for 30 min before cell viability measurement by MTT assay 24 h later. Each data point is the mean of 6 replicated ±SEM.

5.5.2.2 CAG repeat sizing

In the first instance, a panel of LB cell lines, including a 43Q, 44Q and 52Q line, and one from a fast progressing Track-HD subject with the p.R507H variant (see Methods), were cultured and exposed to either 50 or 150 µM H_2O_2. None of the lines with 42-44 CAG repeats showed significant change in modal CAG repeat length or instability index in culture over 17 weeks, or with chronic oxidative stress over 6 weeks. The 52Q line did not show expansion in routine culture, and did not survive treatment with 150 µM H_2O_2, but those exposed to 50 µM may have shown a small expansion, with modal CAG length increasing at a rate of 113.96 ± 6.32 days/Q ($p = 0.0034$) and instability index at a rate of 90.22 ± 5.55 days/Q ($p = 0.0043$), though this was based on an increase of <1Q.

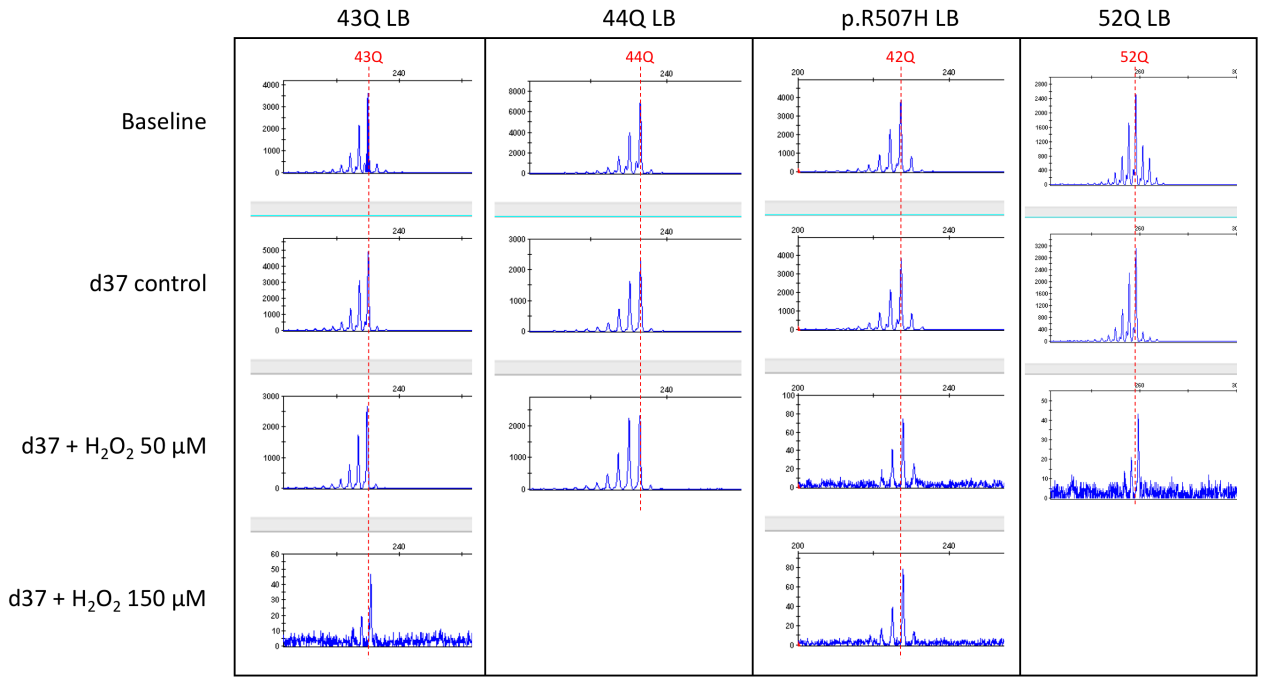


Figure 5.12. Representative CAG repeat sizing in lymphoblastoid (LB) cells chronically stressed with the indicated H_2O_2 concentration.

Modal CAG repeat length at baseline is indicated by a red dotted line. The 44Q and 52Q LB lines exposed to 150 µM H_2O_2 died out after 7d and 2d respectively.

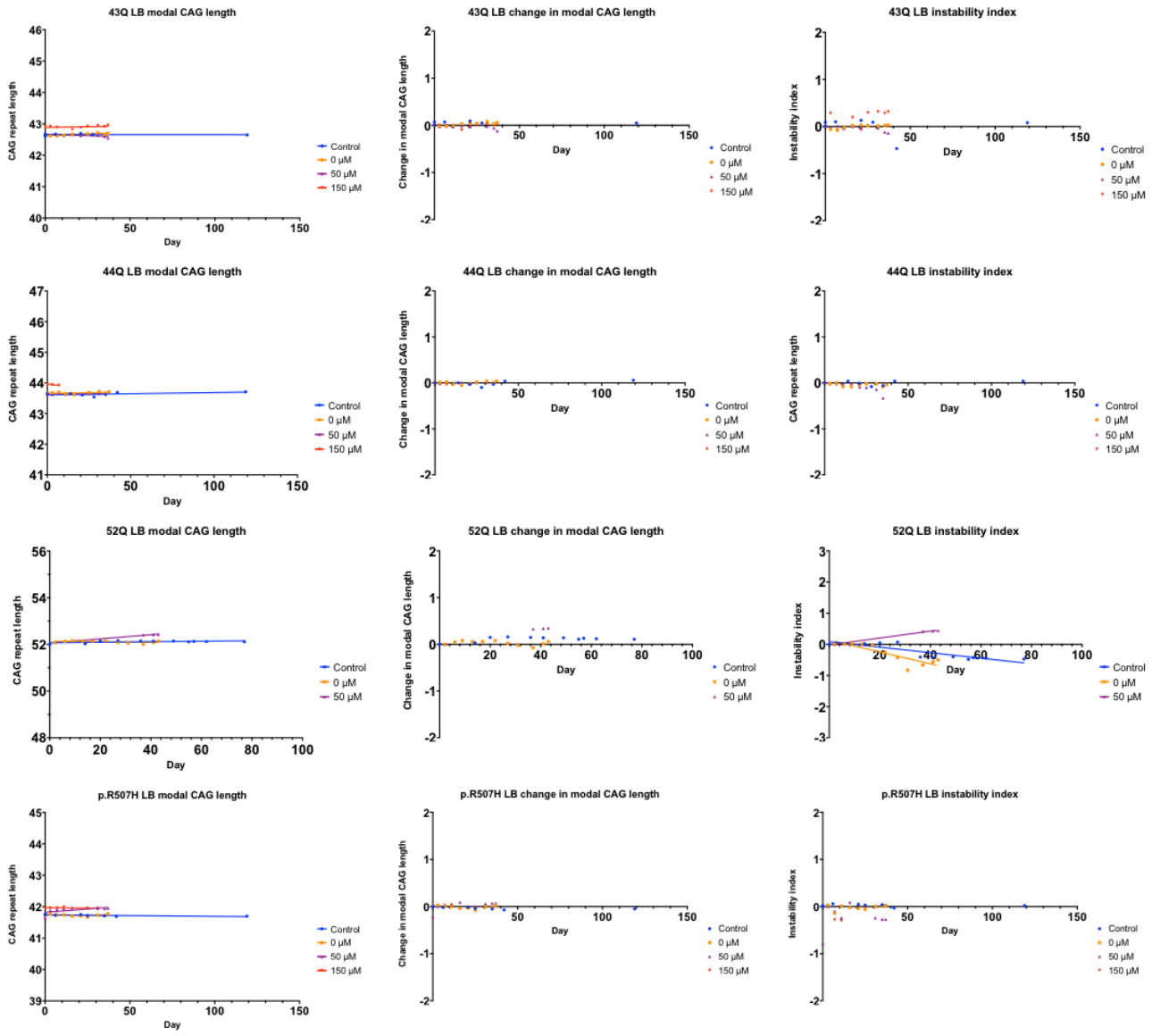


Figure 5.13. Repeat expansion analysis in a 43Q, 44Q, 52Q and p.R507H LB lines chronically exposed to oxidative stress. ***Left column*** *– modal CAG repeat length,* ***middle column*** *– change in modal CAG repeat length,* ***right column*** *– instability index, measured in CAG units.*

5.5.3 250Q lymphoblasts

LBs derived from a juvenile-onset subject, originally estimated to have 250 CAG repeats by PCR and Southern blot, were cultured. The triplet repeat-primed PCR (TP-PCR) capillary electrophoresis trace consistently showed stutter in 3 bp units, with exponential decay as amplification efficiency declines with increasing product size (Jama et al., 2013). This characteristic ladder of stutter peaks is due to the chimeric reverse primer, which is located partially within the CAG region, hybridising to multiple locations within the CAG repeat. As the main allele, at around 250 CAG repeats (Nance et al., 1999), is too large to amplify, it is not possible to accurately assign an allele size by this method. The largest resolved peak was at 136Q, which is the size this allele was called by the UCLH neurogenetics lab.

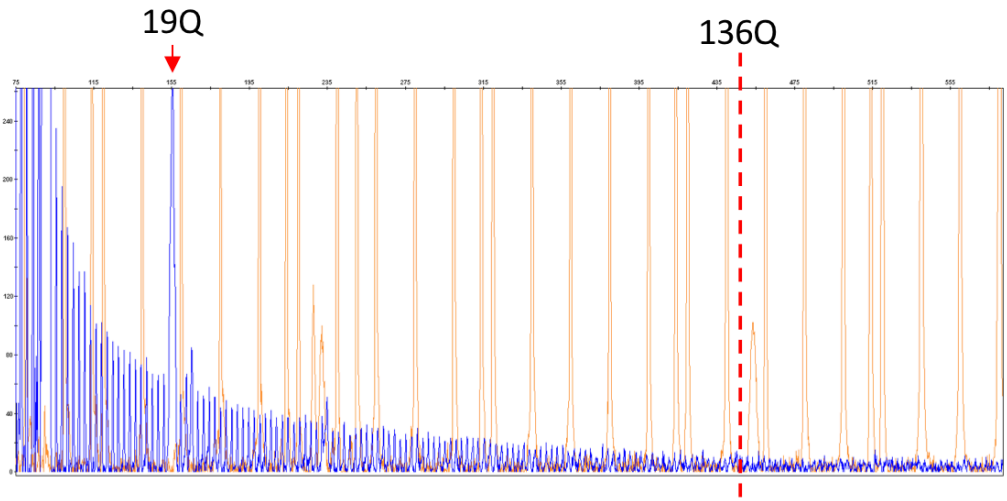


Figure 5.14. CAG repeat sizing in 250Q lymphoblasts (LB) by TP-PCR capillary electrophoresis.

5.5.4 73Q induced pluripotent stem cells

Three clones of induced pluripotent stem cells (iPSC), generated by Sendai reprogramming of fibroblasts from a juvenile-onset HD subject with 73 CAG repeats, were studied for repeat stability.

5.5.4.1 Hydrogen peroxide titration

iPSCs derived from juvenile-onset HD patients with 73 or 109 CAG repeats (QS3.2 and 109Q iPSC respectively) were assayed using the cytotoxicity kit. QS3.2 were more sensitive than 109Q iPSCs, with 50% cell death at around 70 µM and 250 µM H_2O_2 respectively.

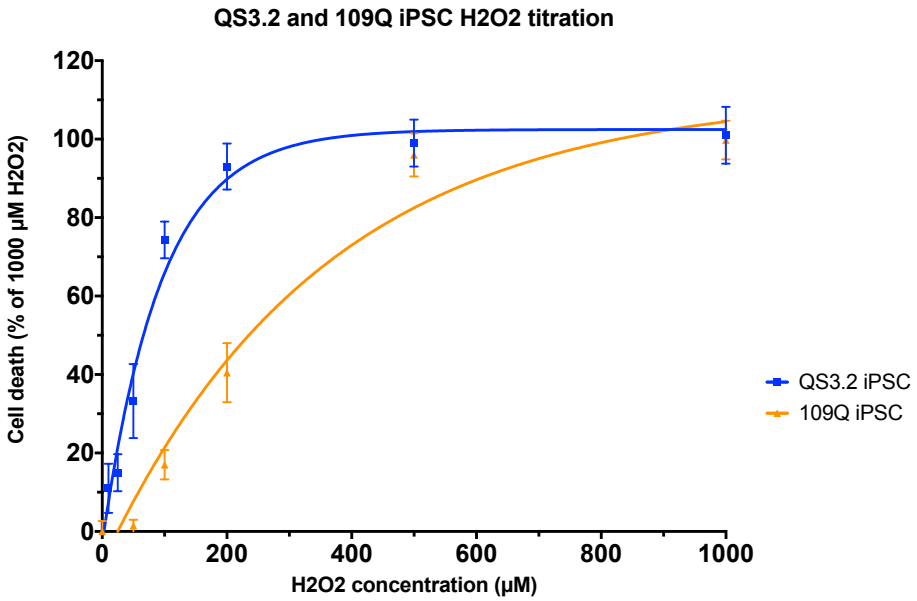


Figure 5.15. Oxidative stress in QS3.2 and 109Q iPSCs.
QS3.2 – clone 2 of iPSCs derived from a juvenile-onset HD subject with 73 CAG repeats. 109Q iPSCs – derived from a juvenile-onset HD subject with 109 CAG repeats.

5.5.4.2 FAN1 knockdown

Expression of *FAN1* was significantly reduced in 73Q QS3.2 iPSCs and 109Q iPSCs by the introduction of shRNA, with transcript levels reduced by 41% and 59% (p = 2.17E-06), and protein by 76% (p = 0.014) and 84% (p = 1.87E-05)

respectively. Transcript levels of *HTT*, *MSH3* and *MLH1* were unchanged (see Figure 5.26). *FAN1* knockdown was maintained throughout differentiation.

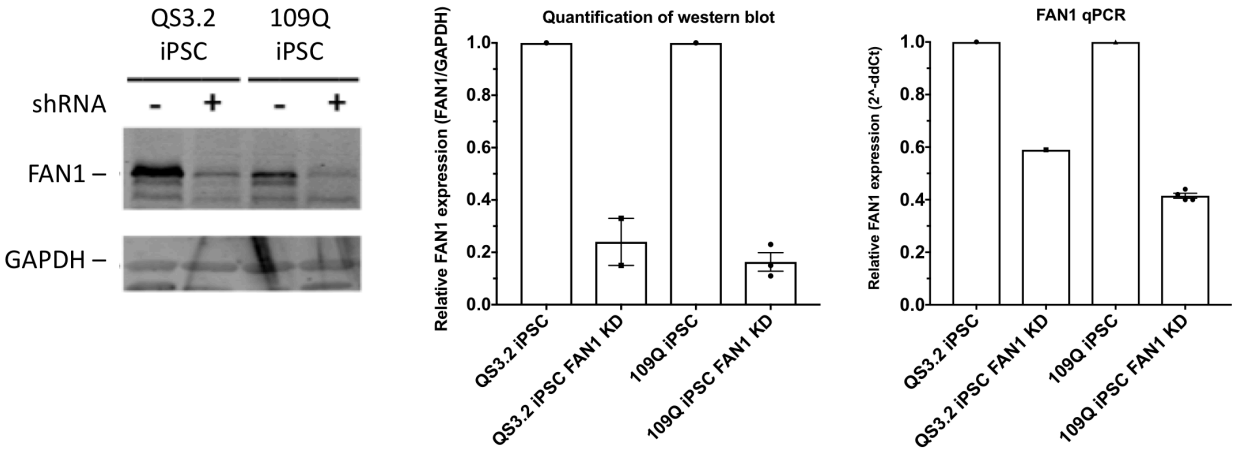
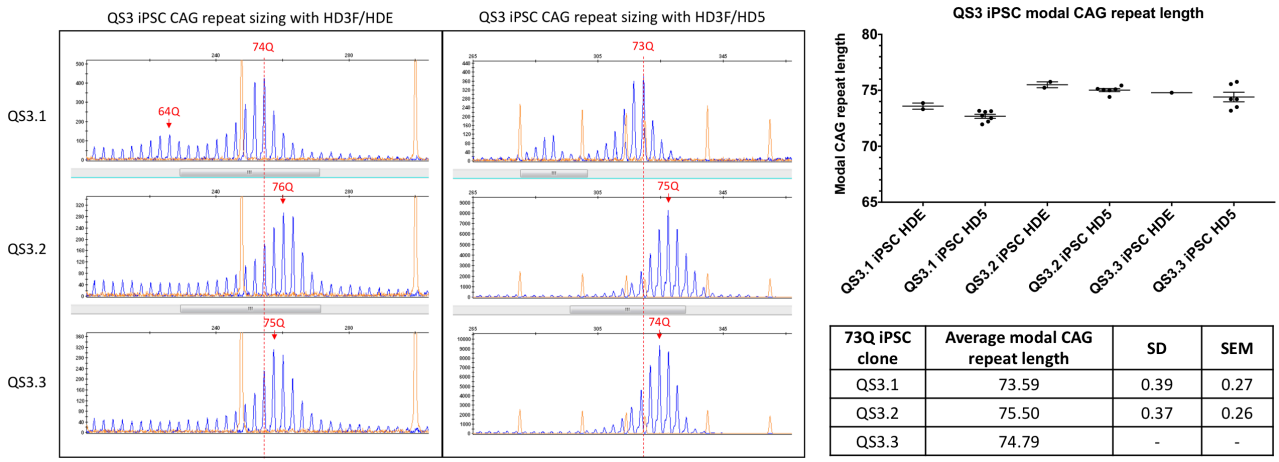


Figure 5.16. shRNA mediated FAN1 knockdown in QS3.2 and 109Q iPSCs.
Left *– western blot.* ***Middle*** *– densitometric quantification of western blot in ImageJ.* ***Right*** *– qPCR of FAN1 relative expression. QS3.2 – clone 2 of iPSCs derived from a juvenile-onset HD subject with 73 CAG repeats. 109Q iPSCs – derived from a juvenile-onset HD subject with 109 CAG repeats.*

5.5.4.3 CAG repeat sizing

At baseline, the CAG length differed between clones, with QS3.1 measuring 74, QS3.2 at 76 and QS3.3 at 75 CAG repeats. Additionally, QS3.1 consistently showed a bimodal distribution, all of which suggest repeat instability during reprogramming.



73Q iPSC clone	Average modal CAG repeat length	SD	SEM
QS3.1	73.59	0.39	0.27
QS3.2	75.50	0.37	0.26
QS3.3	74.79	-	-

Figure 5.17. Baseline CAG repeat sizing in QS3 iPSCs.

Left *– representative CAG repeat sizing in three clones of QS3 iPSCs derived from a juvenile-onset HD patient with 73 CAG repeats. The left panel of this shows triplet-repeat primed PCR (TP-PCR) using the HDE reverse primer for accurate CAG sizing. Note the characteristic ladder of PCR stutter peaks extending from the non-pathogenic allele (out of crop) to the pathogenic allele (shown). The right panel shows CAGCCG PCR using the HD5 reverse primer for optimal amplification. Note increased peak height on the scale relative to TP-PCR.* ***Top right*** *–HDE triplet-repeat primed PCR (TP-PCR) for accurate repeat sizing, HD5 CAGCCG PCR for optimal amplification. Error bars represent SEM.* ***Bottom right*** *–Mean, standard deviation (SD) and standard error of the mean (SEM) are given. Note QS3.3 was sized only once using the HDE primers for TP-PCR.*

Clones 1 and 2 (QS3.1 and 3.2) were cultured under control conditions long term, differentiated to neural stem cells (NSC) or medium spiny neurons, chronically exposed to oxidative stress and *FAN1* was stably knocked down by shRNA. Chronic oxidative stress was induced by exposure to 25 µM or 100 µM H_2O_2 for QS3.2 and 109Q iPSC respectively for 30 min before each passage, aiming to kill around 25% of cells.

QS3.1 and QS3.2 cells cultured in control conditions for 100 days, under chronic oxidative stress for 20 days or differentiated to NSCs and maintained in culture for 43 days did not show any significant change in modal CAG repeat length, instability index or proportional expansion analysis. For iPSCs differentiated into medium spiny neurons (MSNs), the modal CAG repeat number did not significantly increase over 193 days, but instability index increased at a rate of 88.42 ± 13.82 days/Q ($p = 1.01E-03$) and proportional expansion showed a nominal increase (slope = 0.001 ± 5.01E-04, $p = 0.081$). However, this slow rate is equivalent to an increase of only 2 CAG units over the course of the 28 week experiment.

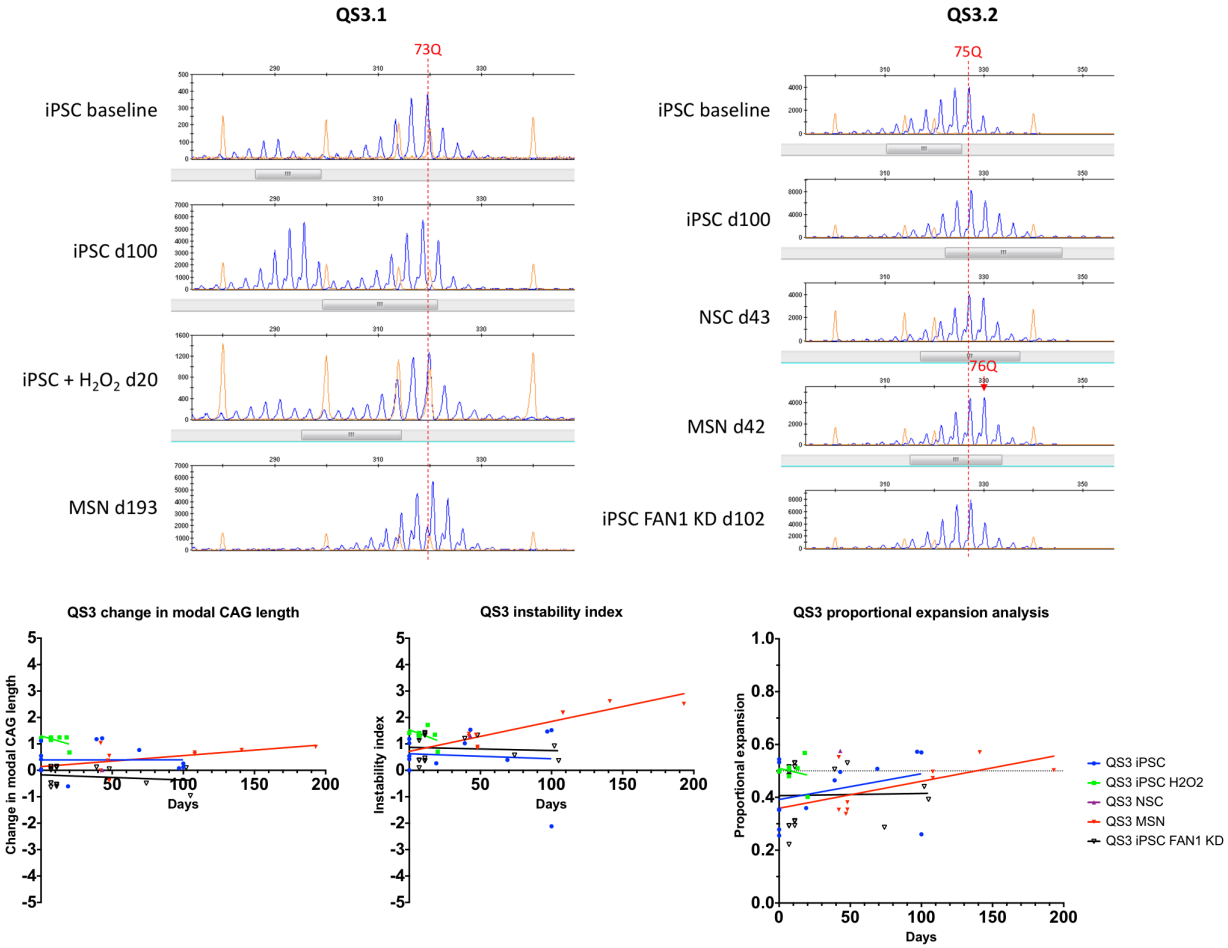


Figure 5.18. CAG expansion analysis in QS3 iPSCs in culture, chronic oxidative stress and differentiation as NSCs or MSNs. ***Top left*** *– representative CAG repeat sizing traces from clone 1.* ***Top right****– clone 2.* ***Bottom*** *– change in modal CAG length from baseline (left), instability index (middle), and proportional expansion analysis (right). QS3 – data from clones 3.1 and 3.2 grouped together for analysis. H_2O_2 – chronic exposure to 25 µM hydrogen peroxide for 30 min before each passage. NSC – iPSCs differentiated along the MSN protocol until passage 2, then maintained as mitotic neural stem cells (NSC). MSN – medium spiny neurons. FAN1 KD – shRNA-mediated stable FAN1 knockdown.*

5.5.5 125Q patient-derived cells

5.5.5.1 Lymphoblastoid cells

Whole blood DNA, LB cells and pluripotent erythroid progenitor cells (EPC) were generated from a juvenile-onset HD patient with 125 CAG repeats. Comparing blood samples in 2014 and 2017, there was no significant change in CAG repeat size, within the ± 3-4 CAG error margin of the assay at this repeat length (Bean and Bayrak-Toydemir, 2014, Losekoot et al., 2013).

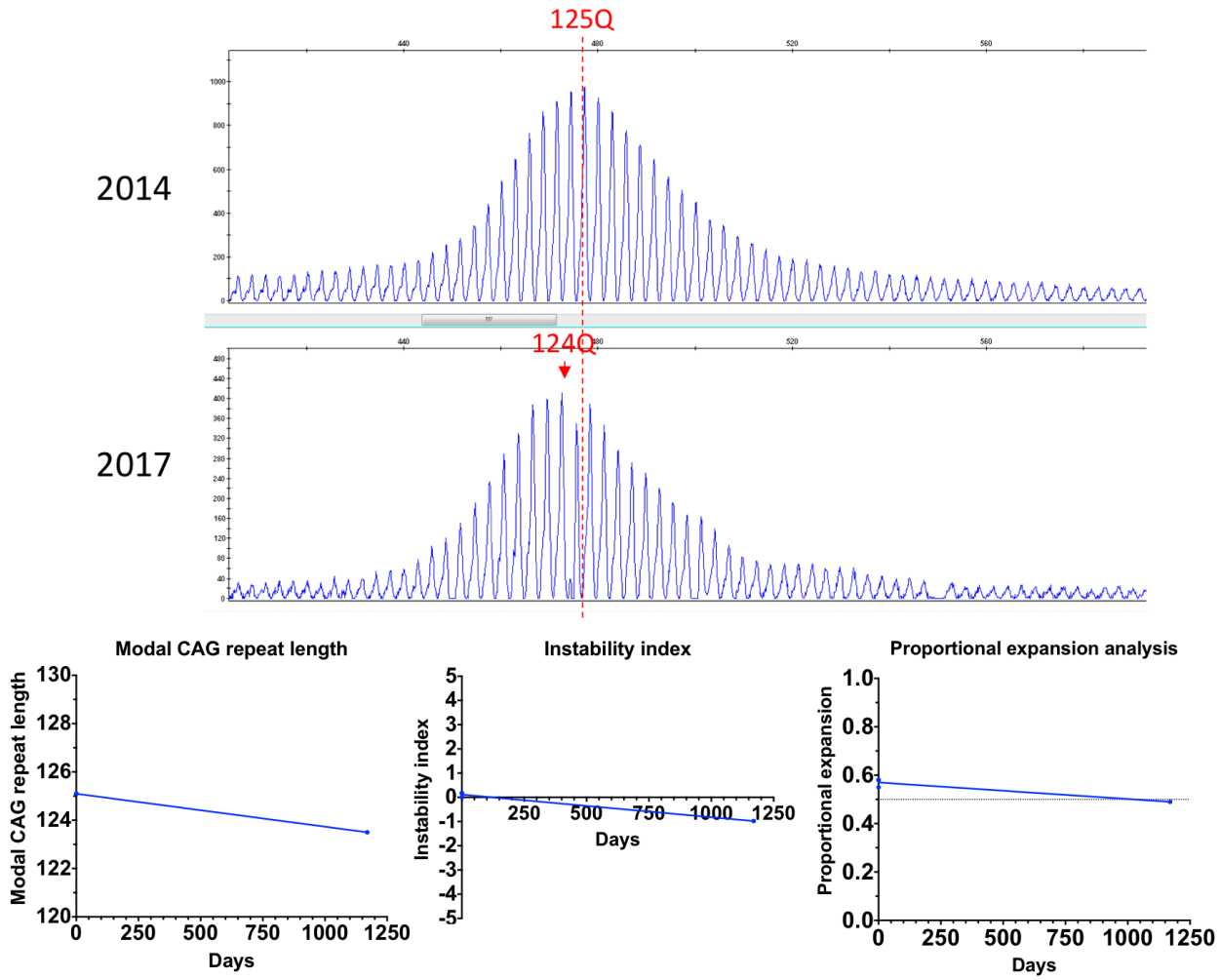


Figure 5.19. CAG repeat sizing of whole blood from a 125Q HD subject sampled 3 years apart.
Top *– representative CAG sizing. Baseline modal CAG repeat length is given by the dotted red line.* ***Bottom left*** *– modal CAG repeat length,* ***middle*** *– instability index,* ***bottom right*** *– proportional expansion analysis.*

At baseline, the LB cells had an average modal CAG repeat length of 136.4 ± 0.58 CAG repeats, meaning the repeat had expanded by 11 CAG during generation of the line. Electrophoresis traces showed a broad distribution of peaks across 72.00 ± 3.86 CAG. In each of five independent cultures the traces became bimodal, before a clonal population arose after around 40 days, as evidenced by the CAG distribution width narrowing to 44.23 ± 5.64 CAG. Consistent with results from single cell cloning in ReN VM NSCs, clones always originated from the higher repeat lengths in the distribution, with an average modal CAG length of 161.2 ± 3.06.

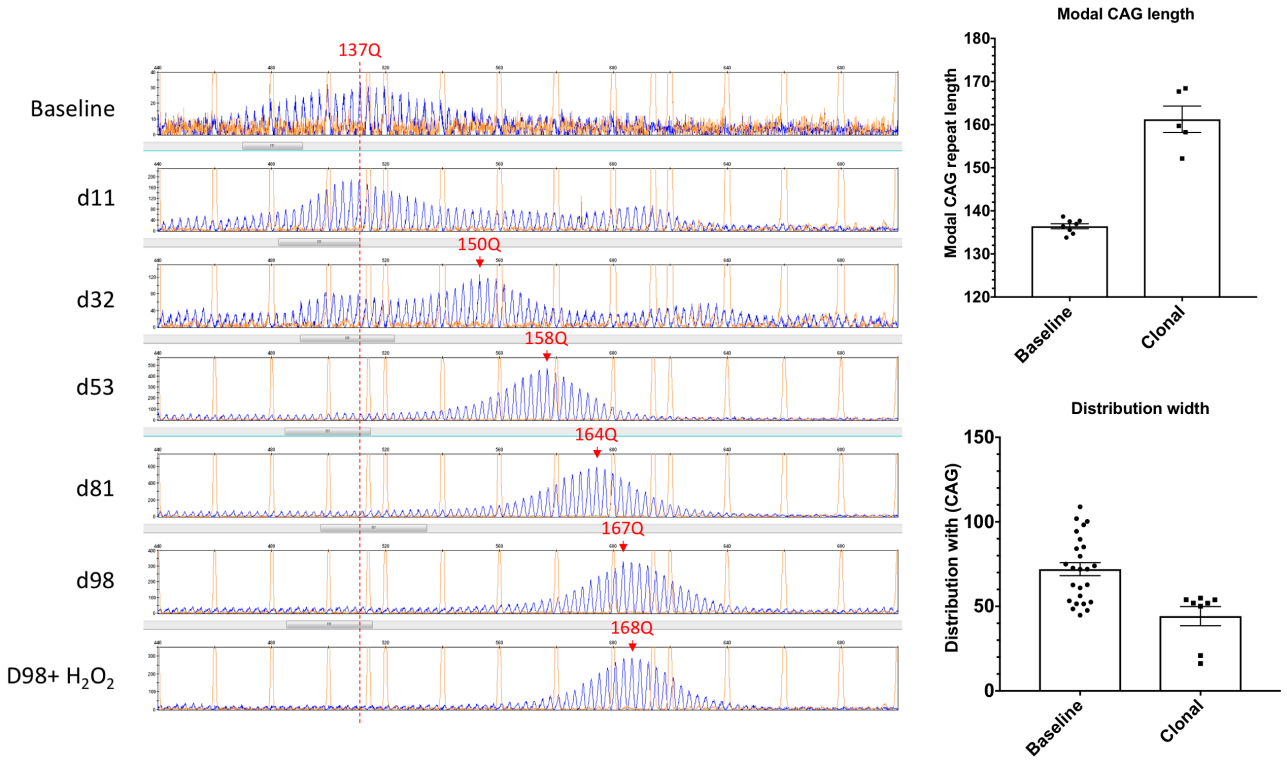


Figure 5.20. CAG repeat sizing of lymphoblasts (LB) from a subject with 125 CAG repeats at baseline and following the emergence of a clone.

Left *– representative CAG repeat sizing in control or chronic oxidative stress conditions. Note the initially broad distribution width at baseline, before a bimodal distribution develops (d11-32) and the emergence of a clonal population after around 40 days. The repeat length in clonal cells then continues to expand. Chronic oxidative stress with 100 µM H_2O_2 for 30 min before every third passage was applied to cells in parallel.* ***Top right*** *– modal CAG repeat length at baseline and in clonal populations after around 40 days.* ***Bottom right*** *– width of the CAG repeat trace on capillary electrophoresis.*

The development of a bimodal distribution confounds expansion analysis by modal CAG repeat length, which appears to suddenly increase up to 30 CAG as the longer repeat becomes dominant. However, the instability index and proportional expansion analysis show a smooth, apparently linear, expansion over time.

Three cultures were exposed to chronic oxidative stress with 100 µM H_2O_2 for 30 min before every third passage, aiming for a sublethal stress that permits long term culture (see Methods). There was no significant change in repeat expansion relative to control conditions. However, stress was applied before clonality was achieved, and cultures would rarely survive longer than 40 days. An acceleration in expansion may be detectable were oxidative stress to be applied after clonality at around day 40.

Focussing on the clonal cells, which increases the sensitivity and resolution of expansion analyses, data was combined from 5 cultures after a clone had emerged. Modal CAG repeat length expanded at a rate of 6.75 $\pm$ 0.66 days/Q (p non-zero = 3.44E-09), proportional expansion analysis increased (p = 8.55E-09), and instability index expanded at a rate of 3.91 $\pm$ 0.27 days/Q (p = 3.07E-12).

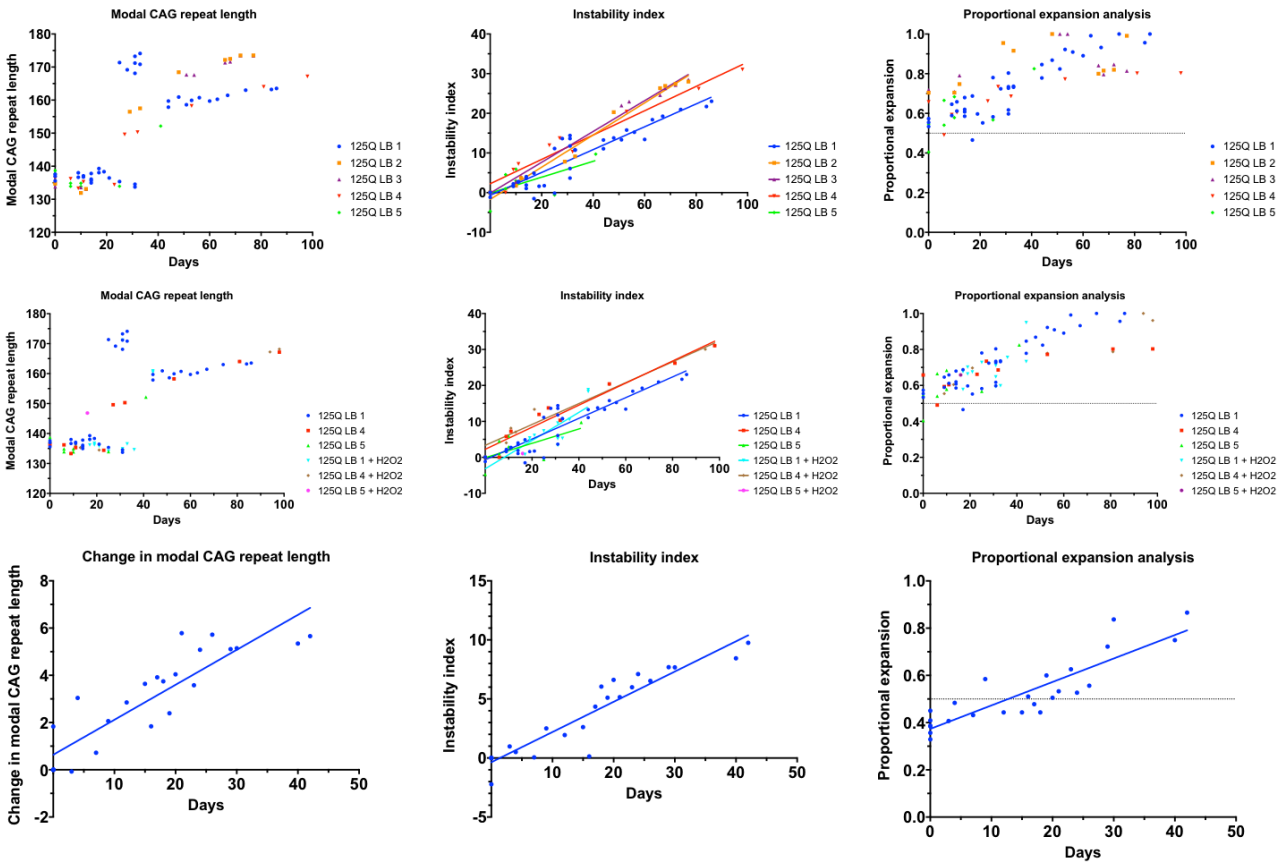


Figure 5.21. CAG repeat expansion analysis in 125 CAG LB cells.

Top row *– 5 independent cultures under control conditions. Note the apparent sudden increase in modal CAG length (top left) around day 30 as cultures become bimodal. Note the linearity of instability index expansion (top middle). On the proportional expansion analysis, expansion is again linear, though this measure reaches its maximum, 1.0, within around 60 days, once the entire distribution becomes larger than the baseline modal CAG length.* ***Middle row*** *– cultured under chronic oxidative stress. 100 µM H_2O_2 for 30 min before every third passage was applied to cells in parallel to cultures 1, 4 and 5. There was no significant difference in expansion between control and oxidative stress conditions.* ***Bottom row*** *– clonal 125Q LB cells. Data from 5 cultures following clonality at around 40 days have been combined.*

5.5.5.2 Induced pluripotent stem cells (iPSC)

Pluripotent erythroid progenitor cells (EPC) generated from the same subject also showed CAG repeat instability on initial characterisation, with modal CAG repeat length expanding at a rate of 25.77 ± 3.82 days/Q (p = 4.57E-03), proportional expansion analysis increasing significantly (p = 0.010) and instability index increasing at a rate of 26.17 ± 4.71 days/Q (p = 0.010). These cells will form the basis of further investigation of the role of DNA maintenance in somatic instability in MSNs.

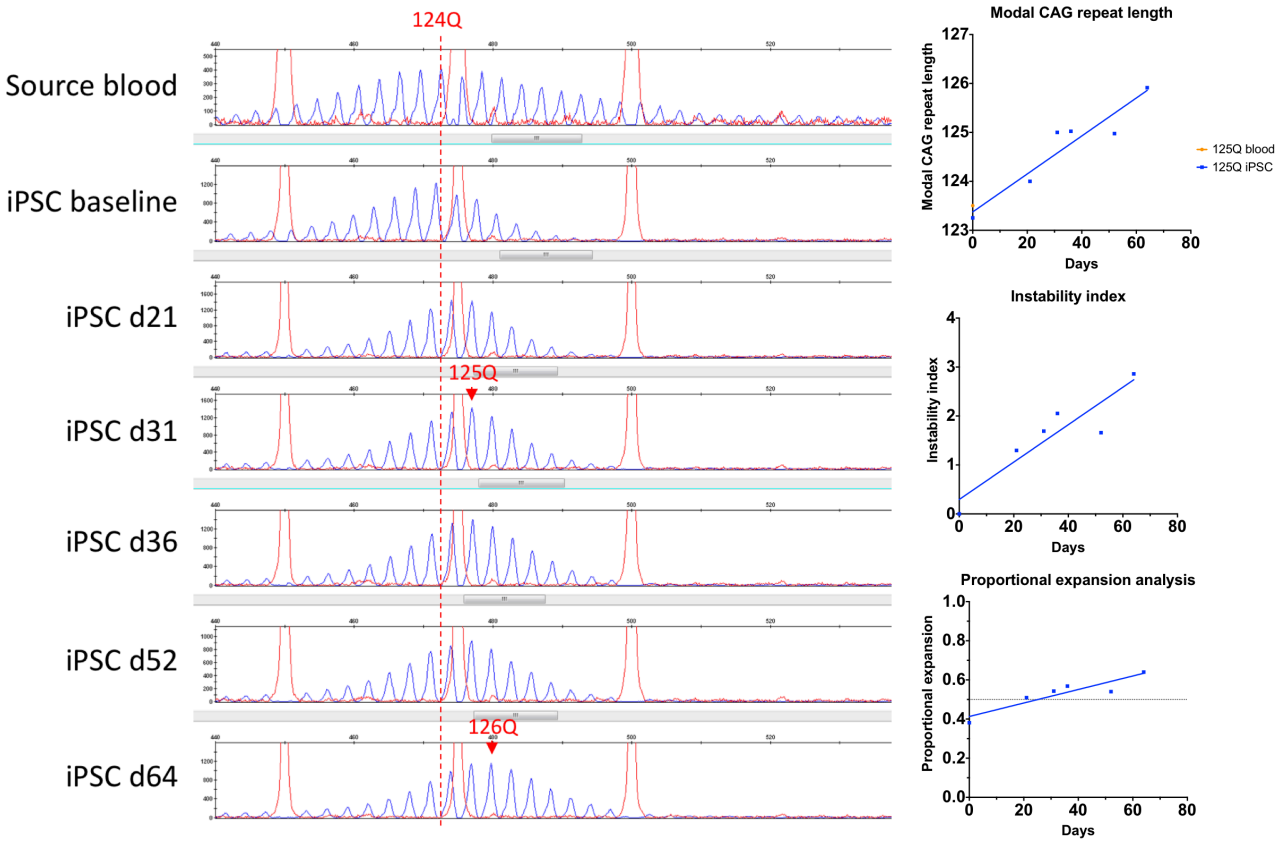


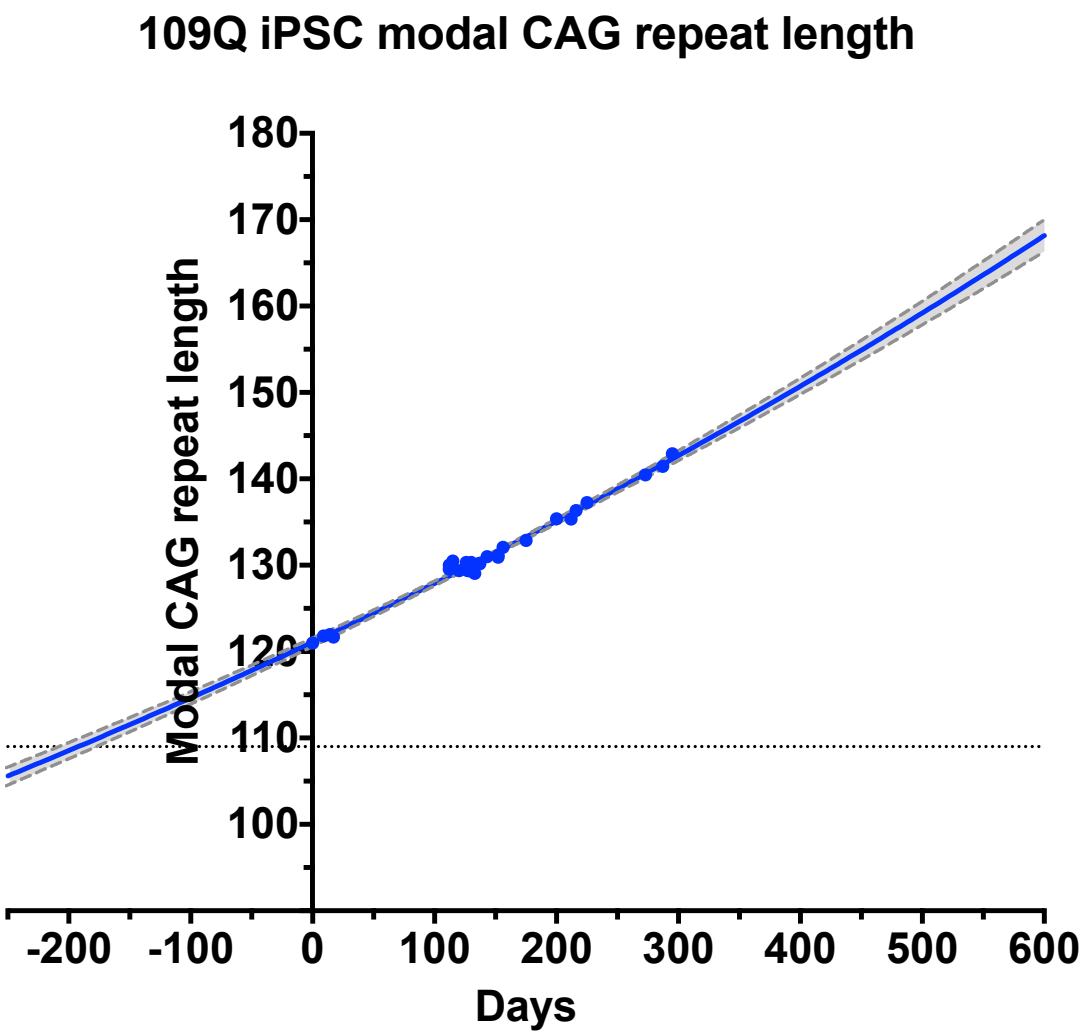
Figure 5.22. CAG repeat sizing in 125Q iPSCs.

DNA extracted from the source blood sample was sized (top), then serial samples were assayed from iPSCs in culture over 64 days. Modal CAG repeat length at baseline is shown by a dotted red line and modal CAG at other timepoints by a red arrow.

5.5.6 109Q induced pluripotent stem cells

5.5.6.1 CAG repeat expansion

iPSCs derived from a juvenile-onset HD patient with 109 CAG repeats were assessed for instability. At baseline, they sized at 121 CAG repeats, indicating expansion from the original length. In three independent cultures the modal CAG repeat length expanded exponentially (r^2 = 0.989, p = 7.65E-26). If it is assumed this curve can be extrapolated to the past, then we observe that the cells have been in culture for 192 days (6 months) since they were derived from the donor 109Q subject. The mean intersection of the exponential functions from three cultures was at 109Q, at an average day -210.



ln(CAG) = (5.473e-04 * d) + 4.797

Figure 5.23. Exponential model of modal CAG repeat expansion in 109Q iPSCs.
Day 0 represents the start of culture, the dotted line at 109 CAG represents the repeat length measured in blood of the subject from which the cells were derived. d – day of culture $r^2 = 0.9882$, $p = 7.651e\text{-}26$.

Culture #	Duration (d)	a	k	r2	p	Intersection			Mean intersect
						1	2	3	
1	295	121.10 ± 0.23	5.47E-04 ± 1.20E-05	0.988	7.65E-26	-	-250.86d, 105.57Q	-209.20d, 108.00Q	-209.69d, 108.83Q
2	196	129.78 ± 0.36	8.23E-04 ± 3.20E-05	0.981	1.56E-12	-250.86d, 105.57Q	-	-169.00d, 112.93Q	
3	64	136.20 ± 0.48	1.11E-03 ± 1.04E-04	0.983	8.66E-03	-209.20d, 108.00Q	-169.00d, 112.93Q	-	

Table 5.6. Exponential modelling of modal CAG expansion in 109Q iPSCs.
Culture 1 – began 16/1/17 on receipt of the cell line, cultured for 295 days. Culture 2 – cells from culture 1 frozen on 22/5/17 (d127, 130Q), thawed on 2/1/18 and cultured for 196 days. Culture 3 – cells from culture 2 frozen on 1/3/18 (d59, 136Q), thawed on 7/5/18 and cultured for 64 days.

5.5.6.2 Oxidative stress

In a pilot experiment, 109Q iPSCs were exposed to chronic oxidative stress in triplicate with 100 µM H_2O_2 for 30 min before each passage, aiming to kill around 25% of cells (see Methods). Over the short 20 day experiment there was no significant difference in expansion rate between control and oxidative stress conditions, but a longer exposure would be required to definitively determine whether there is an effect.

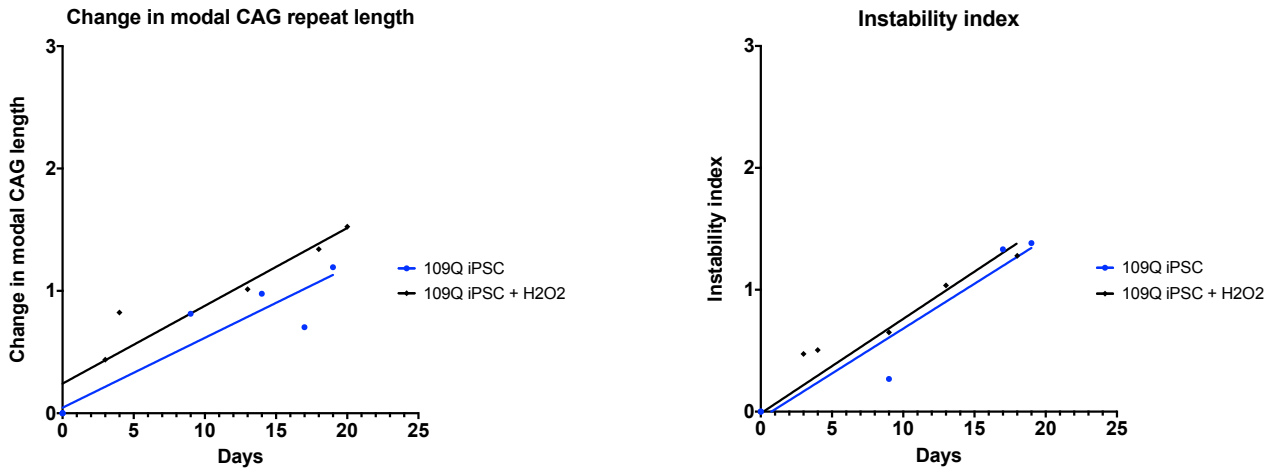


Figure 5.24. CAG repeat expansion in 109Q iPSCs exposed to chronic oxidative stress.
Cells were cultured in control conditions or exposed to 100 μM H_2O_2 for 30 min before each passage over the course of 20 days. Data is generated from duplicate cultures. ***Left*** *– change in modal CAG repeat length,* ***right*** *– instability index.*

5.5.6.3 FAN1 knockdown

Long term, stable *FAN1* knockdown in 109Q iPSCs was achieved by retrovirally mediated shRNA transduction (see Methods). In iPSCs, knockdown at the protein level was 60.26 ± 6.18% relative to empty vector (p = 0.0118) and at the transcript level was 59.34 ± 16.52% relative to control cells (p = 5.67E-03).

An established differentiation protocol, adapted from Arber et al. (2015), was used to generate medium spiny neurons (MSN) from control and retrovirally transduced iPSC lines. qPCR demonstrated significantly increased expression of a panel of mature striatal MSN markers, including genes in the direct (*TAC1*, *DRD1*) and indirect (*PENK*, *DRD2*) striatal pathway, as well as *BCL11B* (*CTIP2*), *OPRM1*, *PENK*, *GAD2* and *CALB1*. Immunofluorescence confocal microscopy demonstrated expression of βIII tubulin (79.56 ± 4.82%), *DARPP32* (13.15 ± 2.19%) and *CTIP2* (31.29 ± 5.23%) in neuronal cells with typical spiny dendritic morphology (Arber et al., 2015).

FAN1 knockdown was maintained throughout differentiation, with protein level reduced by 88.91 ± 11.42% relative to empty vector (p = 0.0112) and transcript level reduced by 50.86 ± 2.96% relative to control (p = 4.68E-03) in MSNs. *FAN1* knockdown did not alter expression of *HTT*, *DARPP-32* or *CTIP2*, or DNA mismatch repair components *MLH1* and *MSH3*.

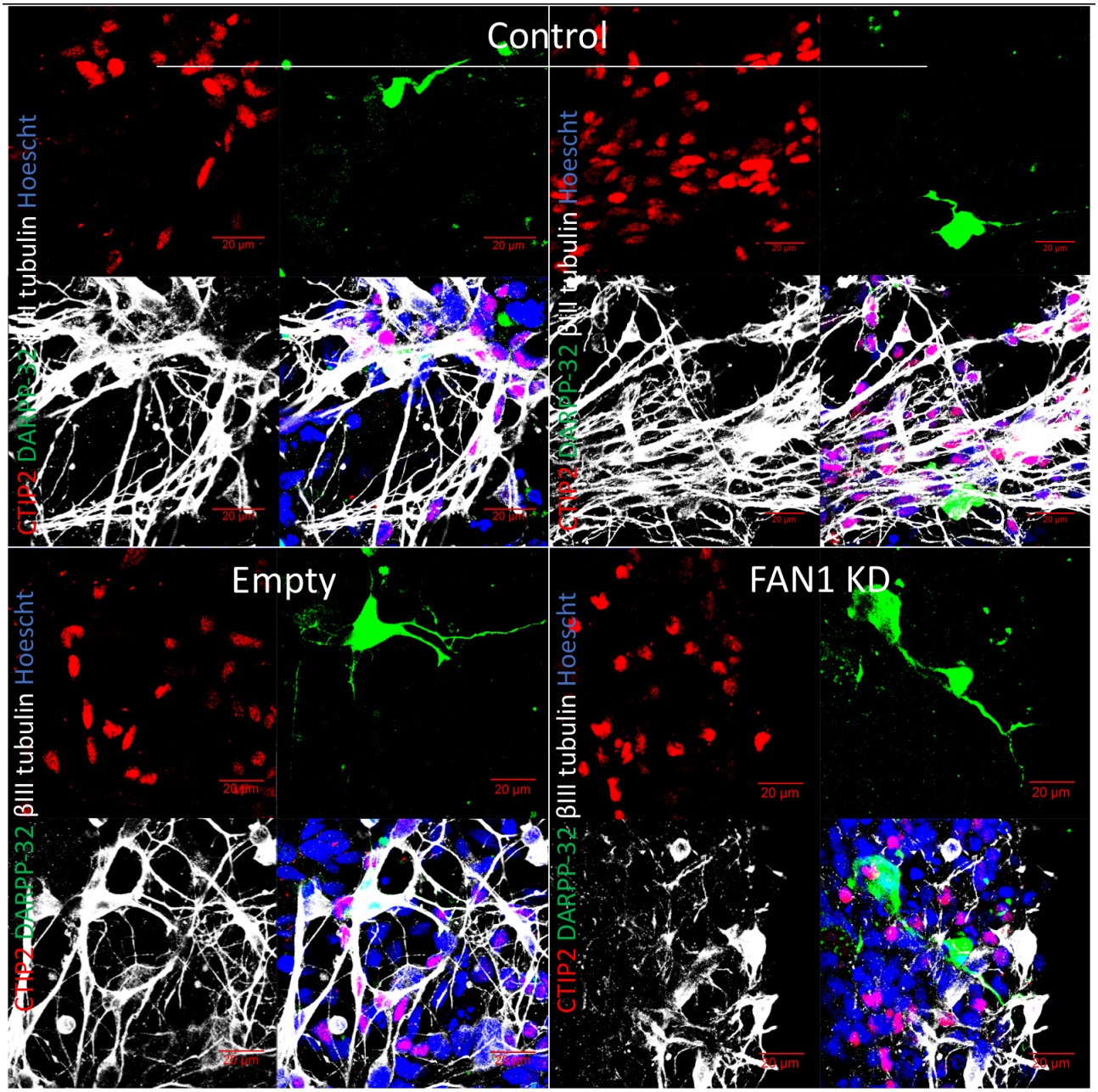


Figure 5.25. Immunofluorescence confocal microscopy of differentiated 109Q medium spiny neurons (MSNs) treated with either FAN1 knockdown, empty vector or in control conditions.

Top row *– control conditions,* ***bottom left*** *– empty vector,* ***bottom right*** *– shRNA-mediated FAN1 knockdown. Primary antibodies are CTIP2 (red), DARPP-32 (green), βIII tubulin (white) and Hoescht (blue). Range bars indicate 20 μM.*

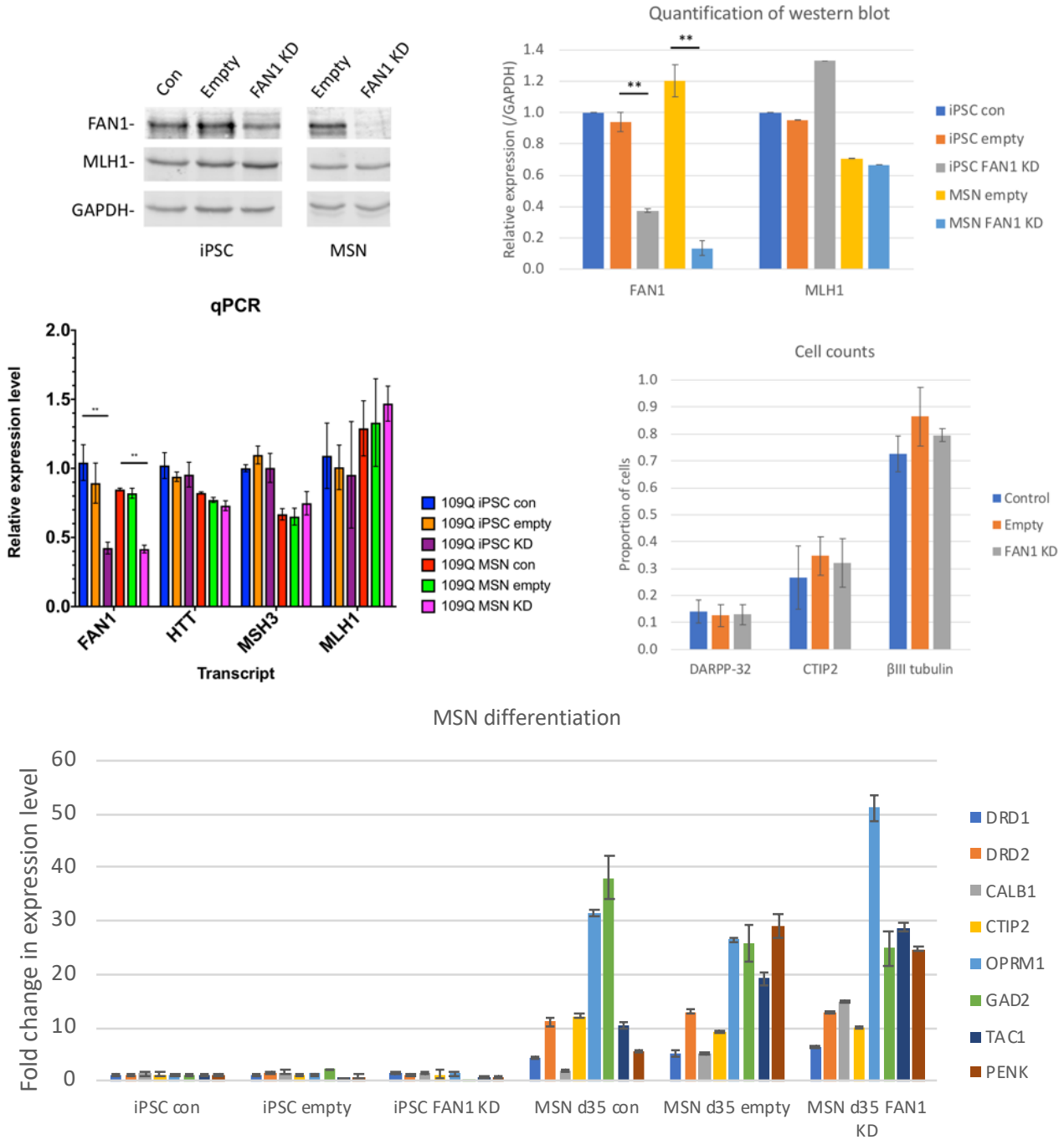


Figure 5.26. Stable shRNA-mediated FAN1 knockdown in 109Q iPSCs and MSNs.
Top left *– western blot of FAN1 and MLH1 in 109Q iPSCs and differentiated MSNs in control, empty vector or FAN1 knockdown conditions.* ***Top right*** *– densitometric quantification of western blot in ImageJ. Data for FAN1 represents the mean of two replicates.* ***Middle left*** *– qPCR of FAN1, HTT, MSH3 and MLH1 in 109Q iPSCs and differentiated MSNs in control, empty vector or FAN1 conditions. Data represents the mean of 5-6 biological replicates for iPSCs and 2 replicates for MSNs. Expression of each gene is relative to the mean of control 109Q iPSCs (blue).* ***Middle right*** *– proportion of cells in MSN cultures positive for DARPP-32 and CTIP2 immunofluorescence staining, expressed relative to Hoescht. Data represents the mean ± SEM of 6 fields for each condition.* ***Bottom*** *– expression of a panel of MSN markers from Straccia et al. (2015). Data represents the mean of 5-6 biological replicates for iPSCs and 2 replicates for MSNs. Expression of each gene is relative to control 109Q iPSCs. Error bars represent SEM. Con – control cells, empty – empty vector, FAN1 KD – shRNA targeting FAN1, MSN d35 – medium spiny neurons at the end of the differentiation protocol.*

In iPSCs, CAG repeat length expanded at a similar rate in control and empty vector treated cells. The rate of change in modal CAG repeat length was 13.95 ± 0.31 days/Q and 16.00 ± 0.72 days/Q respectively (p = 0.383), and instability index

increased at 13.55 ± 0.34 days/Q and 13.33 ± 0.59 days/Q respectively (p = 0.912). Knockdown of *FAN1* increased expansion rate of modal CAG repeat length to 9.81 ± 0.27 days/Q and instability index to 9.50 ± 0.28 days/Q, a significant acceleration relative to control and empty vector treated cells (p mode = 3.15E-15, p index = 4.71E-13 respectively). These results suggest FAN1 protects against expansion of the endogenous *HTT* CAG repeat, at least in mitotic cells.

To assess if this mechanism also operates in non-dividing cells, the CAG repeat was measured in differentiated MSNs. In control and empty vector treated MSNs, it again expanded at a similar rate; modal repeat length increased at 15.46 ± 0.70 days/Q and 17.00 ± 1.99 days/Q respectively (p = 0.430) and instability index expanded at 16.74 ± 0.75 days/Q and 15.01 ± 3.72 days/Q respectively (p = 0.641). The rate was equivalent to iPSCs in control and empty vector conditions (p mode = 0.275 and 0.635). *FAN1* knockdown accelerated expansion of modal CAG length to 10.71 ± 0.77 days/Q and instability index to 9.24 ± 0.56 days/Q, which are significant increases relative to control and empty vector treated cells (p mode = 4.15E-04, p index = 3.28E-08). Once again, expansion rate in *FAN1* knockdown MSNs and iPSCs was equivalent (p mode = 0.357). These results suggest FAN1 restrains CAG repeat expansion in cultures containing a high proportion of post-mitotic differentiated striatal neurons.

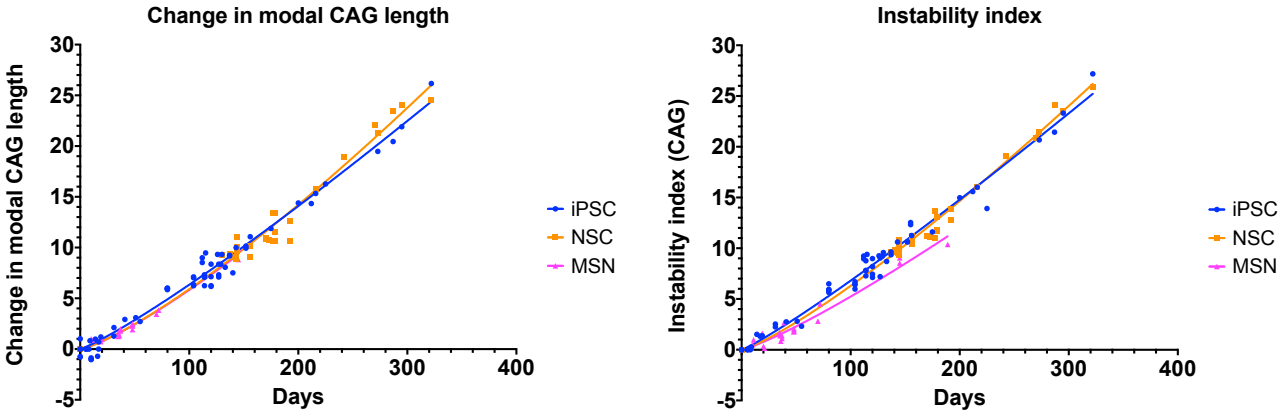


Figure 5.27. Comparison of exponential expansion models in 109Q iPSC, NSC and MSNs.
Representative CAG repeat sizing from a sample culture of iPSCs (blue), neural stem cells (NSC, orange) or differentiated medium spiny neurons (MSN, magenta). ***Left*** *– change in modal CAG repeat length.* ***Right*** *–instability index. Exponential models have been fitted for 109Q iPSC (ln(CAG) = (1.043e-02 * d) + 0.5844; r^2 = 0.7934, p = 2.580e-19), NSC (ln(CAG) = (5.996e-03 * d) + 1.396; r^2 = 0.9434, p = 2.126e-27), and MSN (ln(CAG) = 1.795e-02 * d) – 0.202; r^2 = 0.916, p = 1.106e-09)*

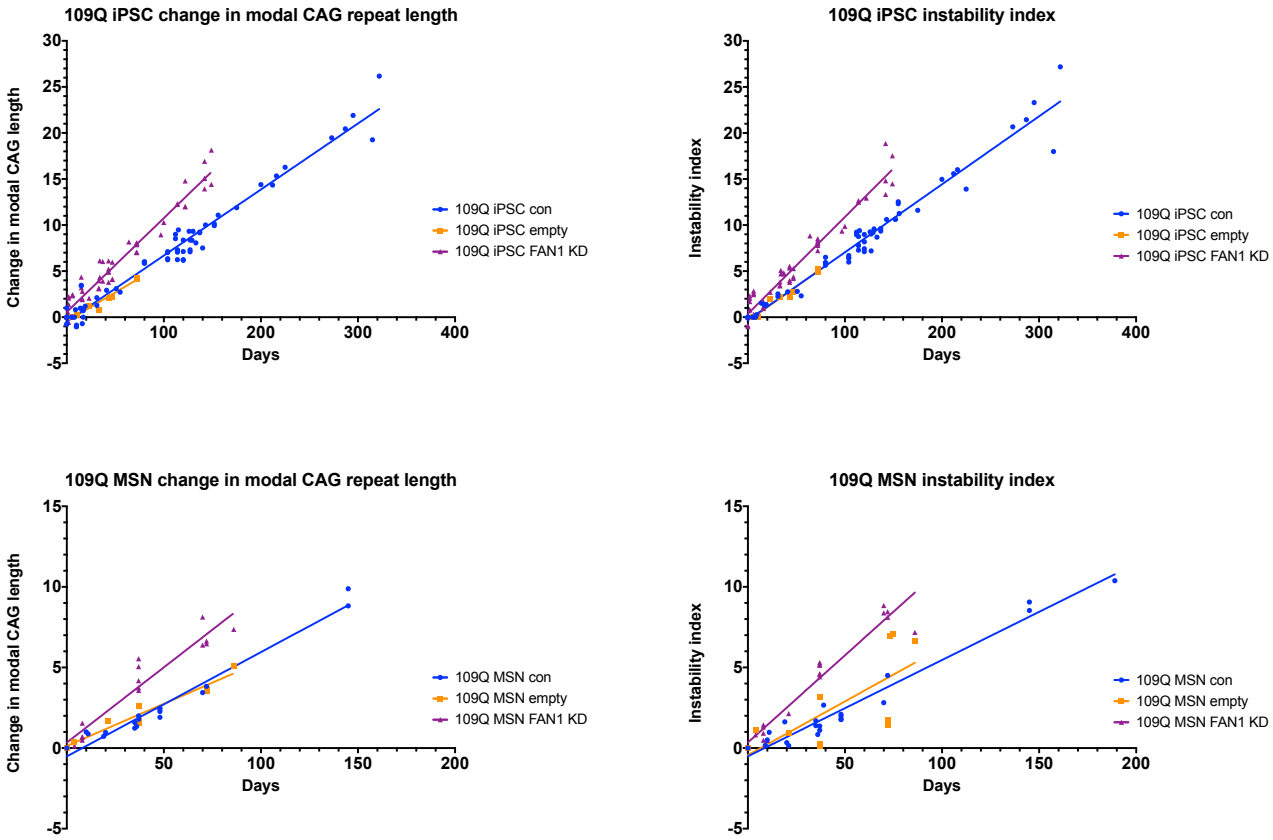


Figure 5.28. CAG repeat expansion in 109Q iPSCs and MSNs following shRNA-mediated FAN1 knockdown.
Linear regression models have been fit. ***Top row*** *– iPSCs,* ***bottom row*** *– differentiated medium spiny neurons (MSN),* ***left column*** *– change in modal CAG repeat length,* ***right column*** *– instability index. Blue – control untreated conditions, orange – empty vector treated, purple – shRNA-mediated FAN1 knockdown. For iPSCs, data is generated from 9 replicates of control conditions, 2-4 replicates from each of two empty vector transductions, and 3-7 replicates from each of three FAN1 shRNA transductions. For MSNs in each condition, data is from 2 replicates of three serial differentiations from each of two independent transductions.*

5.5.7 Comparison of CAG expansion rates

No cell lines with up to 76 CAG repeats showed significant instability in culture. In 109Q iPSCs, expansion was exponential, suggesting expansion rate positively correlates with CAG repeat length. This is supported by lines derived from the juvenile-onset HD subject with 125 CAG repeats, as her iPSCs at baseline expanded at a rate of 26 days/Q, but her clonal LB cells, at 161 repeats, expanded faster than any other cell line, at 6.75 days/Q. Comparing across different cell types the relationship between CAG repeat length and expansion rate was less clear, with 109Q iPSCs expressing 131 repeats expanding at 14 days/Q, but ReN VM neural stem cell single cell clones (SCC) carrying a longer 142 repeat expanding more slowly at 28 days/Q. Expansion rate was not clearly related to mitotic rate; doubling time in iPSCs was approximately half that of LB cells, though their expansion rate was slower, suggesting the mechanism underlying repeat instability is independent of DNA replication.

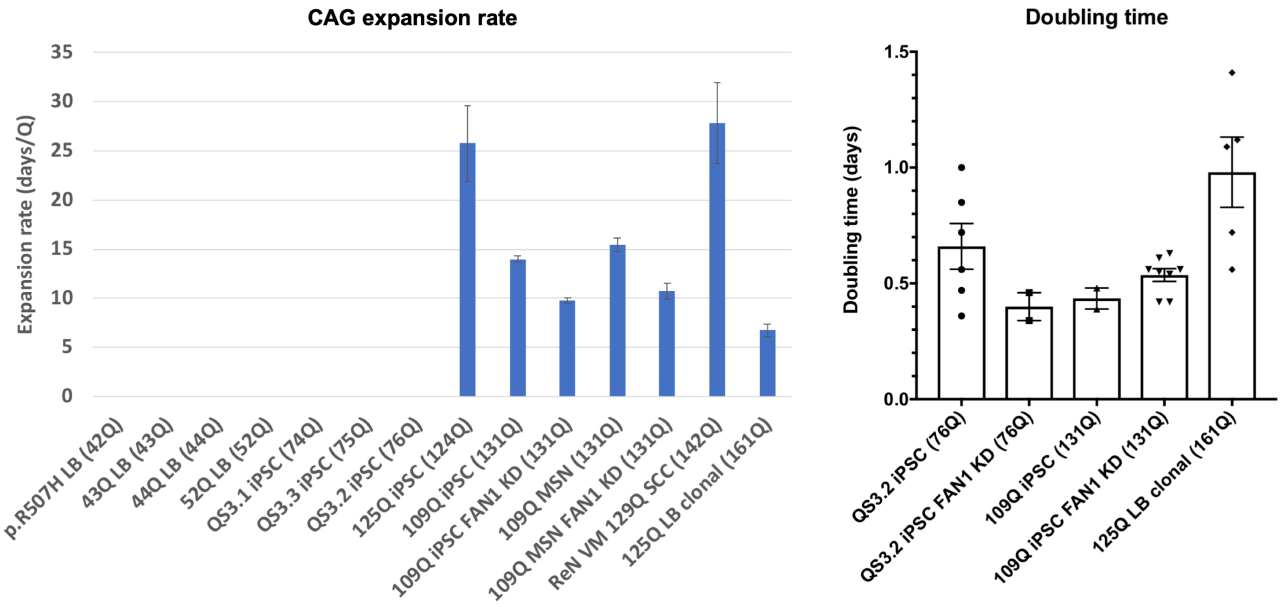


Figure 5.29. Comparison of CAG repeat expansion rate in HD cell lines.

Left *– rate of change in modal CAG repeat length, expressed as days per CAG unit (days/Q). Cell lines sorted in order of CAG repeat length at the point of rate measurement (note this usually differs from original founder repeat length).* ***Right*** *– doubling time of cells in culture. LB – lymphoblastoid cell, p.R507H – FAN1 variant associated with fast disease progression, QS3/109Q/125Q – iPSCs derived from juvenile-onset HD patients with 73, 109 and 125 CAG repeat respectively, ReN VM – ReNeuron neural stem cells, 125Q LB clonal – LBs derived from a juvenile-onset HD patient with 125 CAG repeats following the emergence of a clonal population.*

5.6 Discussion

Pathogenic repeat expansions, such as the CAG in *HTT* exon 1, are unstable somatically and through the germline in a process that is age-dependent, expansion biased and tissue-specific. In HD (Wheeler et al., 1999, Goula et al., 2012, Shelbourne et al., 2007a), DM1 (Anvret et al., 1993, Wong et al., 1995, Ashizawa et al., 1993, Lopez Castel et al., 2011) and Friedreich's ataxia (FRDA) (De Biase et al., 2007, Clark et al., 2007b), expansion occurs in the tissues most prominently affected by the condition. The mechanism driving it remains unclear, but expansion occurs in postmitotic cells and requires DNA repair factors, particularly the mismatch repair pathway, suggesting it happens during DNA repair, rather than replication. Recent genetic studies have identified DNA repair genes *FAN1* (GeM-HD, 2015, Bettencourt et al., 2016) and *MSH3* (Hensman Moss et al., 2017b) as modifiers of HD motor onset and disease progression respectively. Though several artificial and patient-derived cell models of DM1 repeat instability exist, no HD cell models currently show significant repeat expansion within an observable timeframe that would allow measurement of changes in expansion rate induced by manipulation of DNA repair factors. In the present study, 11 human cell lines, including ectopic expression systems and patient-derived lymphoblastoid (LB) and stem cells, were evaluated for robust and quantifiable *HTT* CAG repeat expansion that could form the basis of an investigation of DNA repair.

5.6.1 ReNeuron neural stem cells

ReN VM neural stem cells (NSC) transduced to express *HTT* exon 1 containing either 71 or 129Q did not show significant expansion in routine culture or when exposed to chronic oxidative stress. Though the 129Q repeat in ReN CX cells appeared to expand during differentiation and when exposed to chronic oxidative stress, this may have represented clonal selection rather than true repeat expansion. Their poor retention of *HTT* exon 1 expression through differentiation precluded evaluation of longitudinal change in repeat length.

ReNcell lines expressing 129Q *HTT* exon 1 show a broad, often multimodal distribution of CAG repeat lengths on the capillary electrophoresis trace, which limits the ability to detect repeat expansion. ReN VM 129Q cells were single cell cloned by serial dilution, producing cell lines with significantly narrower, normally distributed traces. Clones tended to arise from cells expressing the longest repeat lengths in the original population, averaging 142Q, suggesting a positive selection pressure for larger alleles. Whilst increasing CAG repeat length produces a protein with a more toxic polyglutamine tract, it reduces protein expression levels (Persichetti et al., 1996) and there may be a threshold repeat length above which expansion is beneficial to cell survival (Dragatsis et al., 2009). Clonal ReN VM cells showed CAG expansion, with modal repeat length increasing at a rate of 27.86 ± 4.12 days/Q ($p = 0.029$).

5.6.2 Track-HD patient-derived lymphoblastoids

Cannella et al. (2009) found lymphoblastoid cells (LB) with at least 64Q showed modest expansion in culture averaging an increase of 3 repeats over 6 months (53.22 ± 16.29 days/Q), and Jonson et al. (2013a) found that chronic oxidative stress can accelerate CAG repeat expansion in transgenic mouse embryonic stem cells. A panel of LB cells from Track-HD patients with repeat lengths ranging from 43 to 52 CAG, and including a fast progressing subject heterozygous for the p.R507H *FAN1* variant associated with early onset (GeM-HD, 2015), showed no significant repeat length change in culture. 52Q LBs appeared to show modest expansion with oxidative stress over 6 weeks, equivalent to a rate of 113.96 ± 6.32 days/Q ($p = 0.0034$). Electrophoresis traces from a 250Q LB line derived from a juvenile-onset HD subject showed characteristic exponentially decaying triplet repeat-primed PCR (TP-PCR) stutter peaks. Whilst these can be used to

diagnose the presence of an expanded allele, this method does not permit accurate sizing and therefore precludes longitudinal repeat expansion analysis.

5.6.3 73Q patient-derived stem cells

Three clones of iPSCs generated from fibroblasts of a juvenile-onset HD patient with 73 CAG repeats were evaluated. At baseline the repeat length of each clone had changed to between 74 and 76 CAG, and one clone, QS3.1, showed a bimodal distribution with a secondary peak at 64Q, all of which suggest repeat instability during reprogramming. However, during routine culture and with chronic oxidative stress, there was no significant change in the CAG repeat length of these iPSCs. With differentiation into medium spiny neurons (MSN) there was a significant increase in instability index ($p = 1.01E-03$), with a trend towards increasing modal CAG repeat length and proportional expansion, though the rate was slow, equivalent to an increase of only 2 CAG repeats in 28 weeks.

5.6.4 109Q and 125Q patient-derived cell lines

iPSCs generated from a juvenile-onset HD subject with 109 CAG repeats were initially sized at 121Q, showing expansion from the original source, and in long term culture the repeat expanded exponentially. LB cells generated from a juvenile-onset HD subject with 125Q showed a broad, multimodal distribution of CAG lengths on repeat sizing. A clonal line arose within 40 days, once again from the longer repeat lengths of the founder population. The clones averaged 161Q, then continued to expand rapidly at a rate of 6.75 ± 0.66 days/Q ($p = 3.44E-09$). Pluripotent erythroid progenitor cells (EPC) generated from the same subject also demonstrated repeat expansion, though at the slower rate of 25.77 ± 3.82 days/Q ($p = 4.57E-03$).

5.6.5 FAN1 protects against CAG repeat expansion

Retroviral transduction of iPSCs with shRNA targeting *FAN1* lowered its transcript level by 60% and protein by 60-80% relative to control cells and those treated with empty vector. A limitation of dividing cell cultures is the opportunity for cell selection effects, which can confound the interpretation of results (Gomes-Pereira et al., 2014a), but using differentiated cultures circumvents this and more accurately reflects the post-mitotic neuronal tissue in which the *HTT* CAG expands and which prominently degenerates *in vivo*. The iPSCs were therefore differentiated into MSNs, showing significantly increased expression of a panel of MSN and neuronal markers by immunofluorescence, qPCR and western blot. *FAN1* knockdown was maintained throughout differentiation, and did not affect the expression of *HTT*, MSN markers or mismatch repair genes, including *MSH3* and *MLH1*.

In 109Q iPSCs, modal CAG expansion rate was equivalent in iPSCs and differentiated MSNs under control conditions, at 13.95 ± 0.31 days/Q and 15.46 ± 0.70 days/Q respectively. *FAN1* knockdown significantly accelerated expansion by the same degree in iPSCs and MSNs, increasing the rate to 9.81 ± 0.27 days/Q and 9.24 ± 0.56 days/Q respectively. This suggests that *FAN1* is protective, stabilising the CAG repeat in both mitotic iPSCs and post-mitotic medium spiny neurons.

FAN1 was knocked down to the same level in the 73Q iPSC line, but still no expansion was observed during the course of the 102 day experiment. Retroviral transduction of 125Q LB cells was unsuccessful, with no cells surviving selection, and the 125Q iPSC line is currently undergoing characterisation. Once baseline repeat expansion and medium spiny neuron differentiation has been fully characterised, the 125Q iPSCs will provide an independent line in which to assess the effect of *FAN1* knockdown on repeat stability.

Collectively, these results show CAG repeat expansion proceeded at the same rate in iPSCs and non-mitotic differentiated MSNs, and was accelerated by knockdown of the nuclease FAN1, favouring DNA repair as the source of repeat instability, rather than DNA replication. Comparing the unstable cell lines, there was no significant correlation between mitotic rate and CAG repeat expansion, also supporting a replication-independent mechanism. Excluding the ReN VM single cell clones, longer repeat lengths were associated with faster expansion rate. 125Q iPSCs expanded at a rate of 26 days/Q, for example, whereas the same patients' LB cells, which have a longer doubling time than other iPSC lines but a repeat length of 161Q, expanded at a rate of 6.75 days/Q. These findings tally with results from DM1 cell models, where cell proliferation rate was not associated with expansion (Gomes-Pereira et al., 2001) and mitotic inhibition did not affect expansion rate (Gomes-Pereira et al., 2014a). It will be interesting to determine whether chemical or genetic arrest of the cell cycle in 109Q and 125Q iPSCs has similarly little effect on CAG repeat instability.

Other DNA repair proteins, particularly those of the mismatch repair pathway such as MutSβ (MSH2-MSH3) (Dragileva et al., 2009, Tome et al., 2013a, Lopez Castel et al., 2010, Manley et al., 1999), MutLα (MLH1-PMS2) (Gomes-Pereira, 2004, Gomes-Pereira et al., 2014b, Pinto et al., 2013b) and MutLγ (MLH1-MLH3) (Pinto et al., 2013a), have been strongly linked to CAG repeat expansion and recently genetic variation in *MSH3* has been shown to influence disease progression in HD patients (Hensman Moss et al., 2017b). A natural progression of this work is to assess the impact of knockdown of each on expansion rate in the cell models developed here, beginning with *MSH3*. One would predict from results in animal studies that inactivation of mismatch repair would reduce expansion (Dragileva et al., 2009, Tome et al., 2013a, Lopez Castel et al., 2010, Manley et al., 1999).

5.7 Summary

In this chapter several cell models of CAG repeat expansion have been generated, including patient-derived LB and iPSCs, and neural stem cells ectopically expressing *HTT* exon 1. These recapitulate the time and repeat length dependent expansion observed *in vivo* (Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009, Mangiarini et al., 1997). Expansion rate appears comparable over the course of this experiment in both mitotic cells and differentiated medium spiny neurons (MSN), the cells showing prominent somatic expansion in HD patients and which selectively degenerate early in the disease course. Knockdown of *FAN1* accelerated repeat expansion in both iPSCs and MSNs, suggesting a common DNA repair mechanism in both dividing and non-dividing cells, in which FAN1 protectively stabilises the repeat.

5.8 Publications relating to this chapter

The work presented in this chapter was published in:

FAN1 modifies Huntington's disease progression by stabilising the expanded HTT CAG repeat. Goold, R.*, **Flower, M.***, Moss, D. H., Medway, C., Wood-Kaczmar, A., Andre, R., Farshim, P., Bates, G. P., Holmans, P., Jones, L. and Tabrizi, S. J. ***Hum Mol Genet***, 2018 Oct 24. doi: 10.1093/hmg/ddy375.

* These authors should be regarded as joint first authors.

Chapter 6 FAN1 activity at *HTT* CAG repeat DNA

6.1 Background

6.1.1 FAN1 is a genetic modifier of Huntington's disease

The *HTT* CAG repeat translates into an expanded polyglutamine tract at the N-terminus of the protein which confers toxicity. Longer CAG repeats cause more severe disease with earlier onset and faster progression (Bates et al., 2015c). CAG repeat tracts are inherently unstable and tend to expand both somatically and intergenerationally through the germline. Expansion results in a longer polyglutamine tract which increases toxicity. Somatic instability likely plays an important role in HD pathogenesis, with CAG repeat length increasing over time in the tissues most affected by HD, particularly the striatum, and correlating with disease onset (Swami et al., 2009). Despite the correlation of repeat length with disease course, onset can still differ by several decades in patients with the same CAG repeat length, as measured in blood (Gusella et al., 2014, Langbehn et al., 2010). CAG repeat length accounts for around 56% of variation in onset (Gusella et al., 2014), but up to half of the remaining variability is heritable and therefore due to genetic differences elsewhere in the genome (Wexler et al., 2004a). Processes that influence onset or progression may offer tractable therapeutic targets.

Recent genome-wide association studies (GWAS) have identified genetic variation that influences HD age at onset (AAO) at a chromosome 15 locus, likely underlain by *FAN1*, with at least two independent signals; one advancing and the other delaying onset (GeM-HD, 2015). The most significant coding SNP, and third most significant of all variants, encodes p.R507H, which was associated with onset 6 years earlier than predicted (GeM-HD, 2015). It produces an amino acid change in the DNA binding domain that is predicted damaging *in silico* (Kumar et al., 2009, Adzhubei et al., 2013). Pathway analysis showed sets of DNA repair genes were associated with disease onset, even when *FAN1* was excluded, suggesting FAN1 may be part of a DNA damage response (DDR) network that modulates HD pathogenesis (GeM-HD, 2015). The chromosome 15 locus was replicated in a GWAS of HD progression (Hensman Moss et al., 2017b), and *FAN1* variants from the HD studies were also shown to influence age at onset in the other polyglutamine diseases too (see Chapter 3) (Bettencourt et al., 2016). Fan1 was recently found to protect against CGG repeat expansion in a mouse model of Fragile X syndrome (Zhao and Usdin, 2018). Similar stabilisation of the *HTT* CAG repeat would reduce somatic expansion and could underlie the association of *FAN1* variation with disease course.

6.1.2 FAN1 function

Functional redundancy is common in the DDR, with components participating in multiple independent pathways (Peng et al., 2014, Zhao et al., 2009). Interaction between mismatch repair (MMR) and interstrand crosslink (ICL) DNA repair pathways has been reported (Peng et al., 2014), with FAN1 capable of partially compensating for loss of EXO1 MMR activity (Desai and Gerson, 2014). Therefore, FAN1 and MMR components may influence HD disease course through a shared mechanism. A stable physical interaction between FAN1 and MutLα components MLH1 and PMS2 further supports this hypothesis (MacKay et al., 2010b). Gel filtration has shown a large proportion of FAN1 co-migrates with MLH1 in a high molecular weight complex that includes PMS2, but not ID complex proteins FANCD2 and FANCI in the absence of crosslinking (MacKay et al., 2010b). This suggests the complex with MutLα plays an important, but as yet unidentified role in FAN1 function. There is significant evidence from mouse models that MMR components are required

for somatic instability (Pinto et al., 2013a, Tome et al., 2013a), and MMR pathways were found to be associated with age at onset in the large GWAS of HD patients (GeM-HD, 2015). MMR components *MSH3* and *MLH1* were recently identified as modifiers of disease progression in HD patients (Hensman Moss et al., 2017b, Lee et al., 2017). As both FAN1 (Zhao and Usdin, 2018) and MMR (Zhao et al., 2015a, Schmidt and Pearson, 2016) regulate repeat stability, interactions between these components suggests they contribute to a common pathway (Zhao et al., 2015b, Zhao and Usdin, 2018, Schmidt and Pearson, 2016).

FAN1 is a DNA endo/exonuclease that was originally identified as a component of the Fanconi anaemia (FA) interstrand crosslink (ICL) repair pathway, though its mutation does not result in Fanconi anaemia (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Smogorzewska et al., 2010a). Rather, mutations have been associated with the recessive renal syndrome karyomegalic interstitial nephritis (KIN) (Zhou et al., 2012, Lachaud et al., 2016b, Thongthip et al., 2016), pancreatic (Smith et al., 2016) and colorectal cancers (Segui et al., 2015b), and autism and schizophrenia (Ionita-Laza et al., 2014). Independent of its ICL repair function, FAN1 is also involved in maintaining genomic stability by regulating recovery of stalled replication forks (Chaudhury et al., 2014, Lachaud et al., 2016a). These functions require FAN1 nuclease activity, which resides in the C-terminal VRR Nuc (viral replication and repair nuclease) domain. It is structure specific, binding branched DNA forms such as 5' flaps and then cutting at every third nucleotide. Its crystal structure has been determined bound to artificial DNA substrates, but DNA binding and its endogenous substrate specificity have not been demonstrated in cell systems. *FAN1* knockout sensitises cells to ICLs and delays the resolution of double strand breaks (DSB) induced during the repair (MacKay et al., 2010a, Kratz et al., 2010b), though its role in repair process remains unknown.

6.2 Aims

Genetic variation in *FAN1* has been linked to disease course in HD patients, and in a fragile X mouse model *Fan1* was shown to protect against CGG repeat expansion. This chapter aims to characterise the effect of *FAN1* genetic variation and expression on DNA repair activity and CAG repeat stability. The synthetic cell systems developed also provide an isogenic background on which to investigate the CAG length-dependence of repeat stability. In synthetic and patient-derived cell lines, experiments look for a novel interaction between FAN1 and *HTT* CAG repeat DNA.

6.3 Methods

6.3.1 FAN1 knockdown in HEK 293 cells

Two siRNA oligonucleotides (Dharmacon) targeted the *FAN1* sequences given below (Liu et al., 2010b). siRNA was transfected into cells at 100 nM using Lipofectamine (Thermo, cat #11668-019) according to manufacturer instructions.

- `siRNA 1: AAACCGTACTTGAGAATGA`
- `siRNA 2: GTAAGGCTCTTTCAACGTA`

6.3.2 Transfection of HEK 293 cells with Myc-tagged FAN1

Full length Myc-tagged FAN1 constructs were designed with six silent mutations at the siRNA target sites that render them resistant to knockdown. These were subcloned into the mammalian expression vector pcDNA3.1(+)hygro and used to transiently and stably transfect HEK293T cells.

```
   1 GGTACCGCCA CCATGGAACA GAAACTGATC TCTGAAGAAG ACCTGATGAT GTCAGAAGGG
  61 AAACCTCCTG ACAAAAAAAG GCCTCGTAGA AGCTTATCAA TCAGCAAGAA TAAGAAAAAA
 121 GCATCTAATT CTATTATTTC GTGTTTTAAC AATGCACCAC CTGCTAAACT TGCCTGCCCC
 181 GTTTGCAGTA AAATGGTGCC TAGATATGAC TTAAACCGGC ACCTTGATGA AATGTGTGCT
 241 AACAATGACT TCGTTCAAGT GGATCCAGGG CAGGTTGGCT TAATAAATTC AAATGTGTCT
 301 ATGGTAGATT TAACCAGTGT TACCTTAGAA GATGTAACAC CTAAGAAGTC ACCACCACCA
 361 AAGACAAATT TAACCCCTGG CCAAAGTGAT TCAGCAAAAA GGGAAGTAAA GCAGAAGATC
 421 AGTCCCTACT TTAAAAGTAA TGATGTGGTG TGCAAAAATC AAGATGAGCT GAGAAATCGT
                                                 R145H          A
 481 AGTGTGAAAG TCATTTGTTT GGGAAGCCTA GCATCTAAAT TGTCCAGAAA ATACGTAAAG
 541 GCTAAAAAAT CAATAGATAA GGATGAAGAA TTTGCCGGTT CTAGTCCACA GAGTTCCAAA
 601 TCCACAGTTG TTAAGAGCCT GATTGATAAC TCTTCAGAAA TTGAGGACGA GGATCAAATT
 661 TTGGAGAACA GTTCTCAAAA AGAAAACGTG TTTAAATGTG ATTCTCTAAA GGAAGAGTGC
 721 ATTCCTGAAC ATATGGTAAG AGGAAGTAAA ATAATGGAAG CCGAAAGCCA AAAGGCTACC
                                   E240K    A
 781 CGGGAATGTG AGAAATCAGC CCTCACCCCT GGATTCTCAG ATAATGCGAT CATGTTATTC
 841 TCACCAGATT TCACTCTTAG GAATACATTA AAGTCTACTT CAGAAGACAG TCTTGTAAAG
 901 CAAGAGTGTA TCAAAGAAGT GGTTGAAAAA CGTGAGGCAT GTCATTGTGA AGAAGTAAAA
 961 ATGACTGTTG CTTCAGAAGC TAAAATACAG CTGTCAGATT CAGAGGCAAA ATCTCATAGT
1021 TCTGCAGATG ATGCTTCTGC ATGGAGTAAC ATCCAAGAGG CTCCTCTGCA GGATGACAGT
1081 TGCTTAAACA ATGATATCCC TCACAGCATT CCTTTGGAGC AGGGGTCAAG CTGCAATGGT
1141 CCTGGTCAAA CAACCGGTCA TCCTTACTAC CTTCGGAGTT TCCTTGTGGT GCTGAAGACA
1201 GTGCTCGAAA ACGAAGATGA TATGTTGCTC TTTGATGAGC AGGAGAAGGG AATTGTAACT
1261 AAATTTTATC AGTTATCAGC TACTGGTCAG AAGTTATATG TCAGACTATT CCAGCGGAAA
1321 TTAAGCTGGA TTAAGATGAC CAAATTAGAG TATGAAGAGA TTGCCTTAGA CTTAACACCT
1381 GTGATTGAAG AATTGACGAA TGCAGGCTTT CTACAGACAG AATCTGAGTT GCAAGAACTC
1441 TCTGAAGTGC TTGAACTCCT TTCTGCTCCT GAACTAAAAT CCCTAGCCAA GACCTTCCAC
1501 TTGGTGAATC CCAATGGACA GAAACAGCAG CTGGTGGACG CCTTTCTCAA ATTGGCCAAA
1561 CAGCGTTCAG TCTGCACTTG GGGCAAGAAT AAGCCTGGAA TTGGTGCAGT GATTTTAAAA
       A  R507H
1621 AGAGCCAAAG CCTTGGCTGG ACAGTCAGTA CGAATCTGTA AAGGCCCCAG GGCTGTGTTT
1681 TCCCGCATCT TGCTACTGTT TTCGTTGACC GACTCAATGG AAGATGAAGA CGCCGCTTGT
1741 GGAGGTCAGG GACAGCTTTC AACAGTCCTG TTGGTCAACC TCGGCCGAAT GGAGTTTCCT
1801 AGTTACACCA TCAATCGGAA AACCCACATC TTCCAAGACA GAGATGATCT TATCAGATAT
1861 GCAGCAGCCA CGCACATGCT GAGTGACATT TCTTCCGCAA TGGCCAATGG GAACTGGGAA
1921 GAAGCTAAGG AGCTCGCTCA GTGTGCAAAA AGGGATTGGA ACAGACTGAA AAACCACCCT
1981 TCTCTGAGAT GCCACGAAGA TTTACCACTC TTCCTGCGGT GTTTCACTGT TGGGTGGATT
2041 TATACAAGGA TTTTGTCTCG GTTTGTGGAA ATACTGCAGA GACTTCACAT GTATGAGGAA
2101 GCCGTCAGAG AACTTGAAAG CCTTTTGTCT CAGAGAATTT ATTGTCCTGA CAGCAGAGGC
2161 CGATGGTGGG ATCGACTGGC CCTTAATTTA CACCAGCACT TGAAGCGCCT GGAACCGACT
2221 ATCAAGTGCA TCACAGAGGG GCTGGCGGAT CCGGAAGTCA GAACGGGACA CCGCCTTTCA
2281 CTGTATCAGC GAGCCGTGCG CCTGCGAGAG TCTCCGAGCT GTAAAAAGTT CAAGCACCTC
2341 TTCCAGCAGC TCCCAGAAAT GGCTGTGCAA GATGTGAAAC ACGTGACCAT CACAGGCAGG
2401 CTGTGCCCAC AGCGTGGGAT GTGCAAGTCT GTGTTTGTGA TGGAGGCCGG GGAGGCCGCT
2461 GACCCCACCA CGGTCCTGTG CTCTGTGGAG GAGCTGGCAC TGGCCCATTA CAGACGCAGC
2521 GGTTTTGACC AGGGGATTCA TGGCGAAGGG TCCACCTTCA GCACCCTGTA TGGCCTCCTC
     Q829H      C
2581 CTGTGGGACA TCATCTTCAT GGATGGGATT CCGGATGTCT TCAGAAACGC CTGTCAGGCA
2641 TTCCCCCTGG ACTTGTGCAC AGACAGCTTC TTCACAAGCA GACGCCCAGC CCTTGAGGCC
2701 AGGCTGCAGC TGATTCATGA TGCCCCCGAG GAGAGCCTGC GGGCCTGGGT GGCAGCCACG
2761 TGGCATGAGC AGGAAGGCAG AGTGGCTTCC CTTGTCAGCT GGGATCGCTT CACGTCTCTT
2821 CAGCAAGCTC AGGATCTTGT CTCCTGCCTG GGGGGCCCTG TGCTCAGTGG TGTGTGCAGG
2881 CACCTGGCTG CTGACTTTCG ACACTGTCGA GGGGGCCTCC CCGACCTGGT GGTGTGGAAC
2941 TCCCAGAGCC GTCACTTTAA GCTGGTGGAA GTTAAAGGCC CCAATGATCG TCTTTCACAT
3001 AAGCAGATGA TCTGGCTGGC TGAACTGCAG AAGCTGGGGG CTGAAGTAGA AGTCTGCCAT
3061 GTGGTTGCAG TTGGAGCTAA GAGCCAAAGC CTTAGCTAAT AACTCGAG
```

Table 6.1. Full length Myc-tagged FAN1 construct.

Four variants associated with fast progression (blue), siRNA target sequences (red) and silent mutations that confer siRNA resistance (grey) are highlighted.

6.3.3 Complementation of U20S cells with FAN1 constructs

U20S FAN1$^{-/-}$ cells were complemented with full length WT or variant *FAN1*, as described in Munoz et al. (2014). Variants were first introduced into the pcDNA5/FRT/TO vector (see Appendix) by site directed mutagenesis (SDM).

6.3.3.1 Site directed mutagenesis

FAN1 mutants were generated by site directed mutagenesis (SDM) of the pcDNA5-GFP-FAN1/FRT/TO vector using the QuickChange XL kit (Agilent, cat #200521) according to manufacturer's instructions. Complimentary primers containing the desired mutation were generated.

p.R145H

- Sense 5'-AAATCAAGATGAGCTGAGAAATCATAGTGTGAAAGTCATTTGTTTGG-3'
- Antisense 5'-CCAAACAAATGACTTTCACACTATGATTTCTCAGCTCATCTTGATTT-3'
- Target TGCAAAAATCAAGATGAGCTGAGAAATC[G>A]TAGTGTGAAAGTCATTTGTTTGGGAAGCC

p.E240K

- Sense 5'-AAGTAAAATAATGGAAGCCAAAAGCCAAAAGGCTACCCG-3'
- Antisense 5'-CGGGTAGCCTTTTGGCTTTTGGCTTCCATTATTTTACTT-3'
- Target AAGAGGAAGTAAAATAATGGAAGCC[G>A]AAAGCCAAAAGGCTACCCGGGAATGTGAGA

p.R377W

- Sense 5'-CGGTCATCCTTACTACCTTTGGAGTTTCCTTGTGG-3'
- Antisense 5'-CCACAAGGAAACTCCAAAGGTAGTAAGGATGACCG-3'
- Target CCTGGTCAAACAACCGGTCATCCTTACTACCTT[C>T]GGAGTTTCCTTGTGGTGCTGAAAACC

p.R507H

- Sense 5'-CAAATTGGCCAAACAGCATTCAGTCTGCACTTGGG-3'
- Antisense 5'-CCCAAGTGCAGACTGAATGCTGTTTGGCCAATTTG-3'
- Target CGCCTTTCTCAAATTGGCCAAACAGC[G>A]TTCAGTCTGCACTTGGGGCAAGAATAAGCCTG

p.D960A (Liu et al., 2010b)

- Sense 5'-GGGCCTCCCCGCCCTGGTGGTGT-3'
- Antisense 5'-ACACCACCAGGGCGGGGAGGCCC-3'
- Target TTTCGACACTGTCGAGGGGGCCTCCCCG[A>C]CCTGGTGGTGTGGAACTCCCAGAGCCGTCAC

p.Q829H

- Sense 5'-CGCAGCGGTTTTGACCACGGGATTCATGGC-3'
- Antisense 5'-gccatgaatcccgtggtcaaaaccgctgcg-3'
- Target TGGCCCATTACAGACGCAGCGGTTTTGACCA[G>C]GGGATTCATGGCGAAGGGTCCACCT

p.P894S

- Sense 5'-CTGATTCATGATGCCTCCGAGGAGAGCCTGC-3'
- Antisense 5'-GCAGGCTCTCCTCGGAGGCATCATGAATCAG-3'
- Target GGCCAGGCTGCAGCTGATTCATGATGCC[C>T]CCGAGGAGAGCCTGCGGGCCTGGGTG

Reactions were set up containing 5 µL of 10x reaction buffer, 10 ng of the vector, 125 ng of each primer, 1 µL of dNTP mix, 3 µL of QuikSolution, made up to 50 µL with water, the 1µL of PfuTurbo DNA polymerase (2.5 U/µl) was added. Cycling conditions were 95°C for 1 min, then 18 cycles of 95°C for 50 sec, 60°C for 50 sec, 68°C for 1 min/kb of plasmid (8 min 30 sec in this case), and then a final extension phase of 68°C for 7 min.

Parental methylated DNA was digested by addition of 0.5 µL DpnI (10 U/µL) and incubation at 37°C for 1 h. The product was run on an agarose gel to confirm amplicon size. Bacterial transformation of the PCR product repairs nicks in the DNA and amplifies the product. 45 µL XL10-Gold ultacompetent cells were aliquoted into a 14 ml tube, 2 µL β-ME was added and the cells incubated on ice for 10 min before 2 µL of the DpnI treated DNA was added. The samples were heat pulsed at 42°C for 30 sec, then incubated on ice for 2 min. 500 µL of 42°C SOC medium (Thermo, cat #15544034) was added, then incubated at 37°C overnight, shaking at 225-250 rpm. The reaction was then transferred to ampicillin agar plates and incubated at 37°C for 24 h.

Colonies were picked and grown up in Luria-Bertani (LB) broth with ampicillin (Thermo, cat #10855001) overnight at 30°C. Colony screening by BamHI digestion identified candidates for sequencing. BamHI cuts the pcDNA5-GFP-FAN1/FRT/TO vector at two sites producing two products. 1 µL of template DNA was added to 0.5 µL of BamHI (NEB, cat #R0136S), 2 µL of CutSmart buffer and 16.5 µL water. It was incubated at 37°C for 1h, then run on an agarose gel.

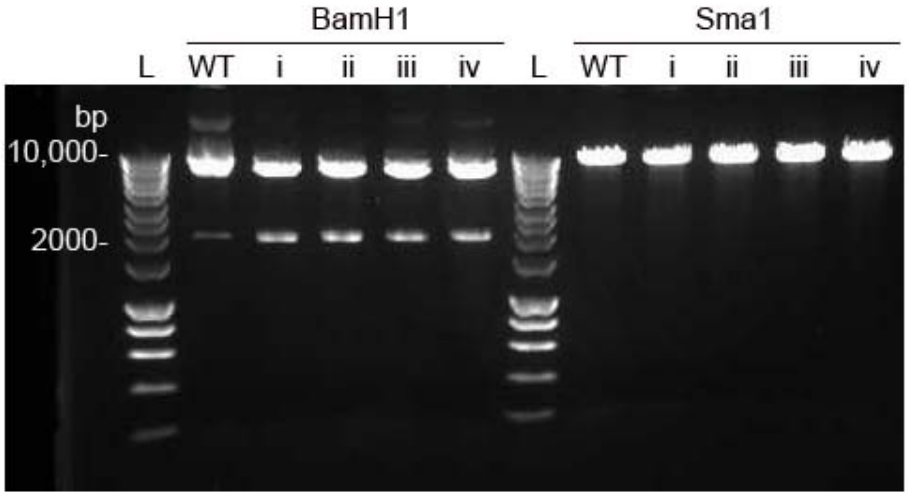


Figure 6.1. Colony screening by restriction digest.
pcDNA5-GFP-FAN1/FRT/TO has two BamHI sites and one SmaI site. WT – wild type vector, i – p.R145H, ii – p.E240K, iii – p.R507H, Iv – p.Q829H.

The plasmids were isolated using the QIAprep spin miniprep kit (Qiagen, cat #27104) according to the manufacturer's instructions. The mutations were confirmed by Sanger sequencing using the following primers.

Primer	Primer sequence	Product (bp)
R145H/E240K sense	CGGACTCGGATCTATGATGTCA	-
R145H antisense	CTGAGAATCCAGGGGTGAGG	782
E240K antisense	CTGGTGAGAATAACATGATCGCA	809
R377W sense	CTCTGCAGGATGACAGTTGC	-
R377W antisense	TCCCTTCTCCTGCTCATCAA	188
R507H sense	TGCAGGCTTTCTACAGACAGA	-
R507H antisense	TGGGGCCTTTACAGATTCGT	269
D960A sense	CACGTCTCTTCAGCAAGCTC	-
D960A antisense	GAAAGACGATCATTGGGGCC	186
Q829H sense	TTCCAGCAGCTCCCAGAAAT	-
Q829H antisense	TCCTGAGCTTGCTGAAGAGA	494
P894S sense	CTGGACTTGTGCACAGACAG	-
P894S antisense	GAGCTTGCTGAAGAGACGTG	186

Table 6.2. Sequencing primers to confirm SDM.

6.3.3.2 Transfection of U20S FAN1$^{-/-}$ cells

First, the Cas9-Flag was flipped out. The reaction mix was made by adding 1 ml Opti-MEM media (Thermo, #31985062) and 60 µL GeneJuice (Millipore, #70967), then vortexing and incubating at room temperature for 10 min (Munoz et al., 2014). 10 µg of pOG44 (MRC PPU, #DU13162) was added, then agitated and incubated for 10 min at room temperature.

The reaction mix was added to cells and incubated for 48h. Zeocin selection was applied at 200 µg/ml (InvivoGen, # 11006-33-0) for 5 days.

The Cas9-Flag flip out cells were then transfected with the pcDNA5/FRT/TO vector encoding *FAN1*. The transfection mix was made by adding 1ml Opti-MEM and 60 µL GeneJuice, which was then vortexed and incubated at room temperature for 10 min. 9 µg of pOG44 and 1 µg of vector were added, then agitated and incubated for 10 min at room temperature. The transfection mix was added to the cells which were incubated for 48h. Then selection with puromycin (Thermo, #A1113802) at 1.5 µg/ml was applied. All variants were then confirmed by Sanger sequencing.

6.3.4 Lentiviral transduction of U20S cells with *HTT* exon 1

HTT exon 1 constructs with 30, 70, 97 or 118 CAG repeats were stably introduced into U20S cells by lentiviral transduction. p'HRsincpptUCOE+htt IRES eGFP human *HTT* exon 1 lentiviral plasmids were previously described in Trager et al. (2014). For transient expression, cells were transfected directly using GeneJuice, according to manufacturer instructions (Merck, cat #70967). For stable integration, it was packaged in lentiviral particles. HEK 293T cells were transfected with packaging vectors and p'HRsincpptUCOE+htt IRES eGFP using Lipofectamine LTX (Thermo, cat #15338500). After 16 h, media was changed, then at 48 h media containing mature lentivirus was harvested. This was filtered and either used directly or frozen at -80°C. U20S media was mixed 1:1 with lentiviral media, supplemented with 8 µg/ml polybrene and added to U20S cells for 24 h. Media was then changed and tetracycline added. For repeat expansion experiments, this was considered the start of the time course. Once at 80-90% confluence, cells were passaged using TrypZean (Lonza, cat #BE02034E). At each passage excess cells were washed by centrifuging at 1000 xg for 1 min, resuspending in PBS, then centrifuged again at 1000 xg for 1 min. PBS was removed and the semi-dry pellet was stored at -80°C ahead of DNA extraction.

6.3.5 Antibodies

FAN1

- S420C sheep polyclonal antibody to human full length FAN1 (MRC PPU)(MacKay et al., 2010b)
- FS2 sheep polyclonal antibody to human full length FAN1 (MRC PPU, CHDI)
- Ab68572 mouse polyclonal antibody to human full length FAN1 (Abcam)
- Ab95171 rabbit polyclonal antibody to a synthetic peptide corresponding to the N-terminal 50 amino acids of human FAN1 (Abcam)

Myc

- PB11 Myc-Tag mouse monoclonal antibody against a synthetic peptide corresponding to residues 410-419 of human c-Myc (Cell signaling, #2276).

FANCD2

- Ab2187-50 rabbit polyclonal antibody to N-terminal fusion protein fragment of human FANCD2 (Abcam)(MacKay et al., 2010b).
- NB100-182 rabbit polyclonal antibody to N-terminal fusion protein of human FANCD2 (Novus), used for immunofluorescence(MacKay et al., 2010b).

MLH1

- Anti-MLH1 mouse monoclonal antibody to full length human MLH1 (BD Pharminigen, #554073)

HTT

- 4C9 monoclonal antibody raised against the proline-rich region (amino acids 65-84) of human *HTT* (Landles et al., 2010, Weiss et al., 2012).

- 2B7 mouse monoclonal antibody raised against the N-terminal portion of human *HTT* (Weiss et al., 2012).

GAPDH

- Rabbit polyclonal antibody raised against amino acids 1-335 representing full length GAPDH of human origin (Santa Cruz, #sc-47724)

GFP

- Rabbit polyclonal raised against amino acids 1-238 representing full length GFP (Santa Cruz, #sc-8334)

γ-H2Ax

- A300-081A rabbit polyclonal antibody

6.4 Contributions

U20S *FAN1* knockout cells were provided by Prof John Rouse of the MRC Protein Phosphorylation Unit, University of Dundee. HEK 293 *FAN1* knockdown and transfection with FAN1 constructs, site directed mutagenesis, complementation of U20S cells with *FAN1* variant forms, cell culture and DNA repair assays were conducted by Michael Flower under the supervision of Rob Goold. *HTT* exon 1 constructs were generated as detailed in Trager et al. (2014), transfected into U20S cells by Rob Goold, cultured by Michael Flower and Rob Goold, and CAG repeat sizing and analysis were conducted by Michael Flower. Immunoprecipitation was conducted by Rob Goold and qPCR by Michael Flower. The work presented in this chapter has been published in Goold et al. (2018).

6.5 Results

6.5.1 FAN1 interstrand crosslink repair function

6.5.1.1 HEK 293 cells

6.5.1.1.1 Transfection with Myc-tagged FAN1

Generating clones stably expressing *FAN1* constructs proved difficult, with most colonies surviving selection being false positives. Only lines expressing wild type and p.R507H *FAN1* were established. Similar results were obtained in HeLa and SH-SY5Y cells (results not shown). Parallel experiments using a Myc-PSMB5 construct in pcDNA3.1, the same vector, generated plenty of positive clones suggesting that exogenous overexpression of *FAN1* may be toxic.

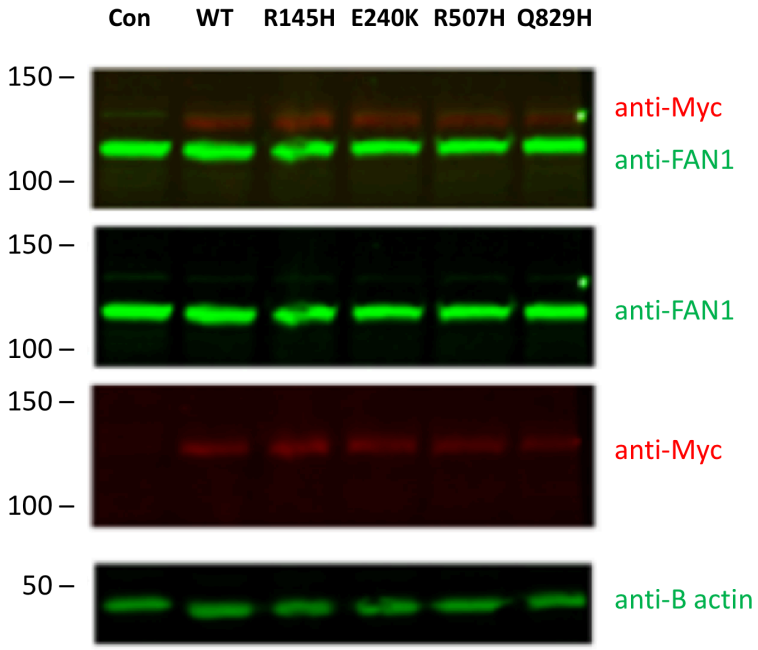


Figure 6.2. Transient transfection of HEK293 cells with full length Myc-tagged FAN1 variants.
Blots are probed with antibodies as shown. Anti-FAN1 antibody S420C, Con – control, WT – wild type. Molecular weights are given in KDa. B actin was used as the loading control

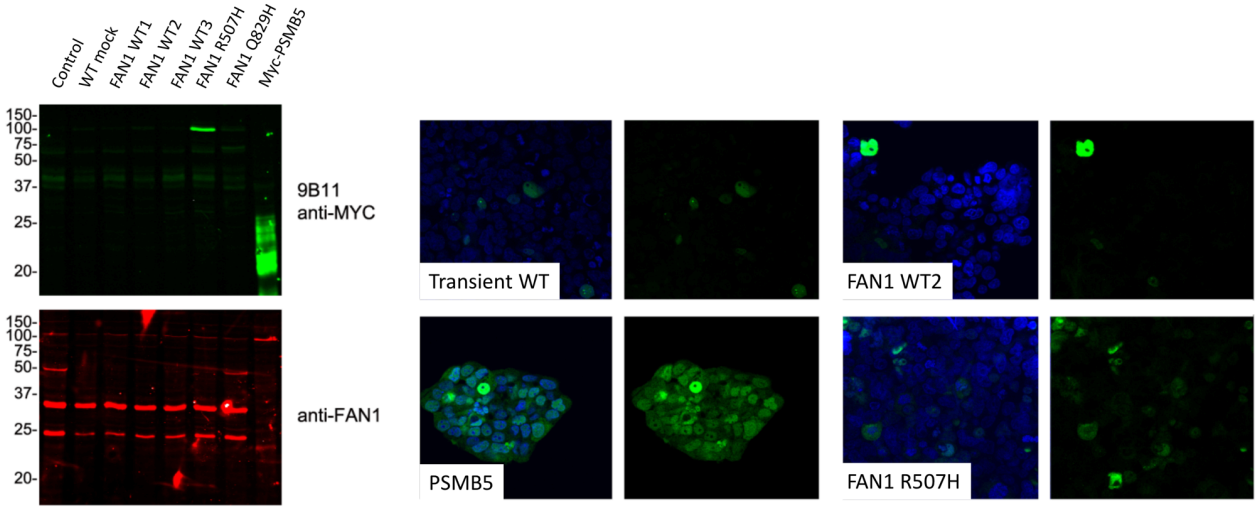


Figure 6.3. Stable transfection of HEK293 cells.

Left *– immunoblot of HEK293 single cell clones transfected with FAN1 variants. Few clones were found to express significant levels of FAN1. Control transfection with Myc-PSMB5 in the same vector provided good expression.* ***Right*** *– confocal immunofluorescence shows HEK 293 cells expressing Myc-tagged constructs. Transient transfection with FAN1 WT (upper left) and stable transfection with FAN1 WT (upper right), Myc-PSMB5 vector control (lower left) or FAN1 p.R507H (lower right). Green – PB11 anti-Myc; Blue – DAPI nuclear stain.*

6.5.1.1.2 Interstrand crosslink repair

HEK 293 cells stably transfected with Myc-tagged wild type and p.R507H *FAN1* formed normal nuclear DNA repair foci that colocalise with FANCD2 after induction of interstrand crosslink (ICL) lesions by MMC.

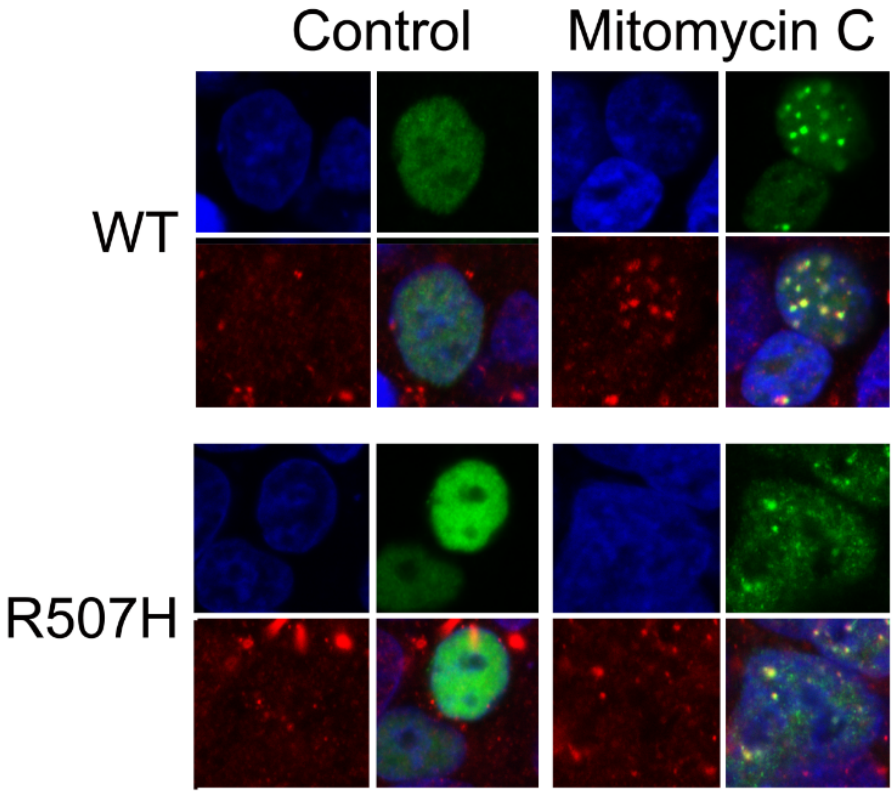


Figure 6.4. Confocal immunofluorescence shows Myc-tagged FAN1 is expressed in the nucleus and forms repair foci that colocalise with FANCD2 following MMC.

Green – PB11 anti-Myc, Red – Anti-FANCD2, Blue – DAPI nuclear stain.

6.5.1.1.3 FAN1 knockdown

Two siRNA oligonucleotides were used to knockdown endogenous *FAN1* in the HEK293 cells. Transfection with either gave around 60% *FAN1* knockdown relative to control. *FAN1* knockdown increased sensitivity to MMC-induced interstrand crosslinks (ICL). Expression of the Myc-tagged constructs was unaffected due to silent mutations at the siRNA target sequences.

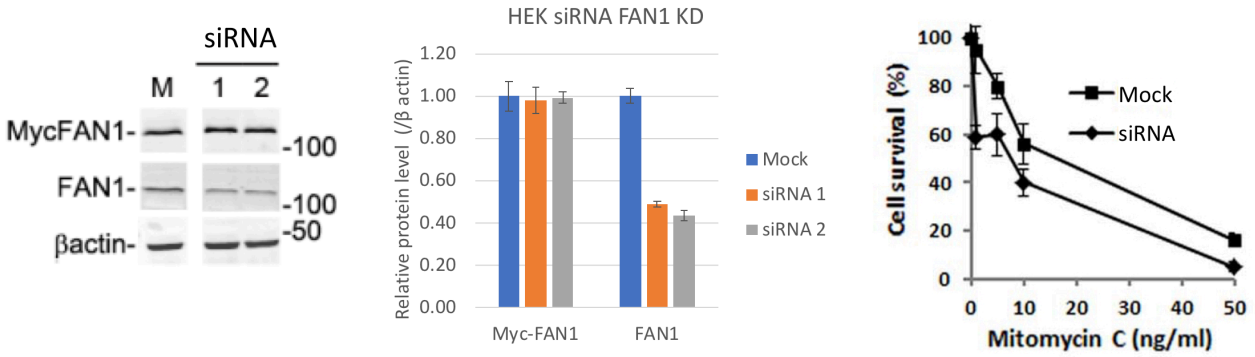


Figure 6.5. siRNA mediated knockdown of endogenous FAN1 in HEK293 cells stably transfected with Myc-tagged p.R507H FAN1. *HEK293 were exposed to either oligonucleotide 1, 2 or mock (M). Cells were harvested after 24h and lysates prepared.* ***Left*** *– representative blot probed with 9B11 anti-Myc and S420C anti-FAN1 antibodies. B actin was used as a loading control. Myc-tagged FAN1 constructs are resistant to knockdown (top panel).* ***Middle*** *– quantification of immunoblots. Endogenous FAN1 levels were reduced to by 53.8 ±2.5% (middle panel).* ***Right*** *– siRNA mediated knockdown of FAN1 in HEK293 cells stably transfected with wild type FAN1 increases sensitivity to MMC in an MTT survival assay. Cells were exposed to MMC at the indicated concentration and survival was assayed at 24h using MTT. Values are the mean of three knockdown experiments, error bars represent SEM.*

6.5.1.2 Lymphoblastoid cells

6.5.1.2.1 Interstrand crosslink repair

Patient-derived lymphoblasts from fast progressing Track-HD subjects with coding variants in *FAN1*, along with their matched HD positive controls, were exposed to MMC to assess resistance to ICLs. Two lines, p.R507H and WT1 (control for p.R145H), showed increased resistance. However, these two lines also expressed *FAN1* at a higher level, which may account for their apparent MMC resistance.

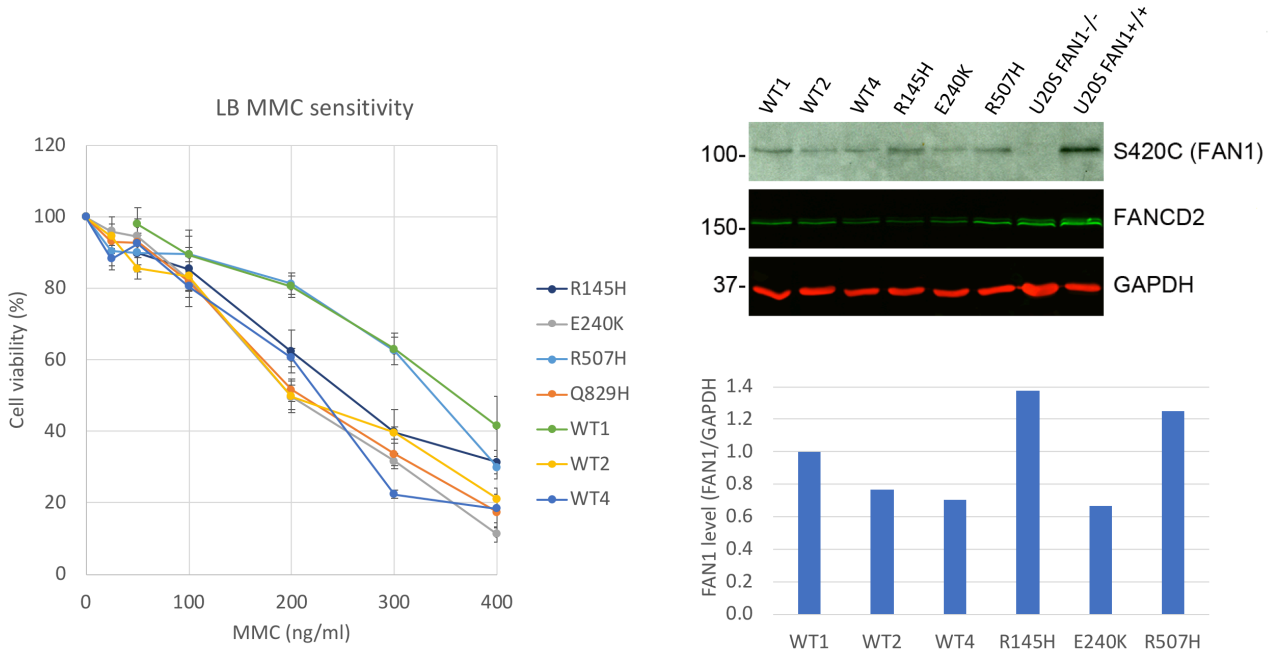


Figure 6.6. Patient-derived LB cell MMC sensitivity.

Left *– LB resistance to MMC as determined by MTT assay at 7-10 days. p.R507H and the WT1 lines demonstrated increased resistance (n=3, ± SD).* ***Right*** *– Immunoblot to quantify FAN1 expression level in HD LB cells. Lysates were prepared from p.R145H, p.E240K and p.R507H subjects and their matched controls. The lower panel shows the blot probed with the S420C anti-FAN1 antibody. GAPDH was used as a loading control. In the lower panel the blot was quantified by densitometry in ImageJ and represented graphically as a ratio of FAN1/GAPDH.*

6.5.1.2.2 Effect of rs3512 on FAN1 expression

To assess whether the rs3512 *FAN1* variant, the most significant SNP from Bettencourt et al. (2016) which was also associated with delayed onset in the GeM GWAS (B = 1.31 years/minor allele, p = 5.28E-13), affects *FAN1* expression level, western blots were prepared from patient-derived HD lymphoblasts of Track-HD subjects. LB cell lines were cultured from three homozygous variant (GG), two heterozygotes and six homozygous wild type (CC) Track-HD patients. Note, the genotyping assay for this variant was designed in reverse, meaning the wild type allele is C and the variant is G. In this small cohort, rs3512 genotype did not significantly affect *FAN1* expression level.

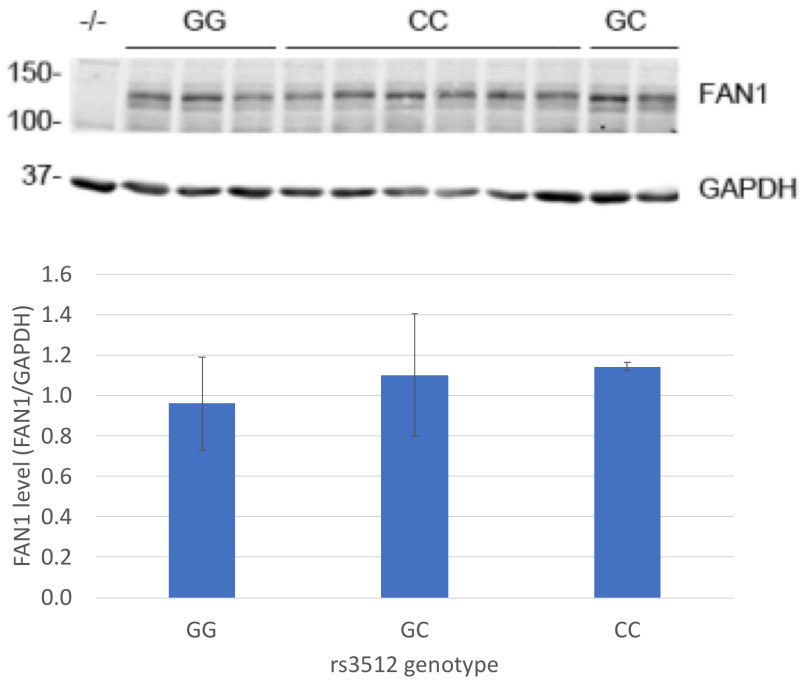


Figure 6.7. Immunoblot for FAN1 expression in HD lymphoblasts from subjects with the given rs3512 genotype.
As the reverse strand was genotyped, C is ancestral and G is variant. Samples were loaded with progression rank decreasing from left to right in each genotype (faster to the right). The upper panel shows the blot probed with S420C FAN1 antibody.

6.5.1.3 U20S cells

6.5.1.3.1 Site directed mutagenesis to introduce FAN1 variants

FAN1 variants were introduced into the pcDNA5-GFP-FAN1/FRT/TO vector by site directed mutagenesis (SDM), then these were used to complement U20S Flp-In FAN1$^{-/-}$ cells, selected using puromycin, and Sanger sequencing confirmed successful transfection. These U20S cells stably expressing the variants permit the assay of GFP-tagged FAN1 function on an isogenic background. p.R507H was the *FAN1* coding variant most significantly associated with HD motor onset in the GeM-HD GWAS (GeM-HD, 2015), p.R145H, p.E240K and p.829H are *FAN1* coding variants found in fast progressing Track-HD patients, and p.D960A is a mutation in the active site that completely ablates nuclease activity (Kratz et al., 2010a, Liu et al., 2010b).

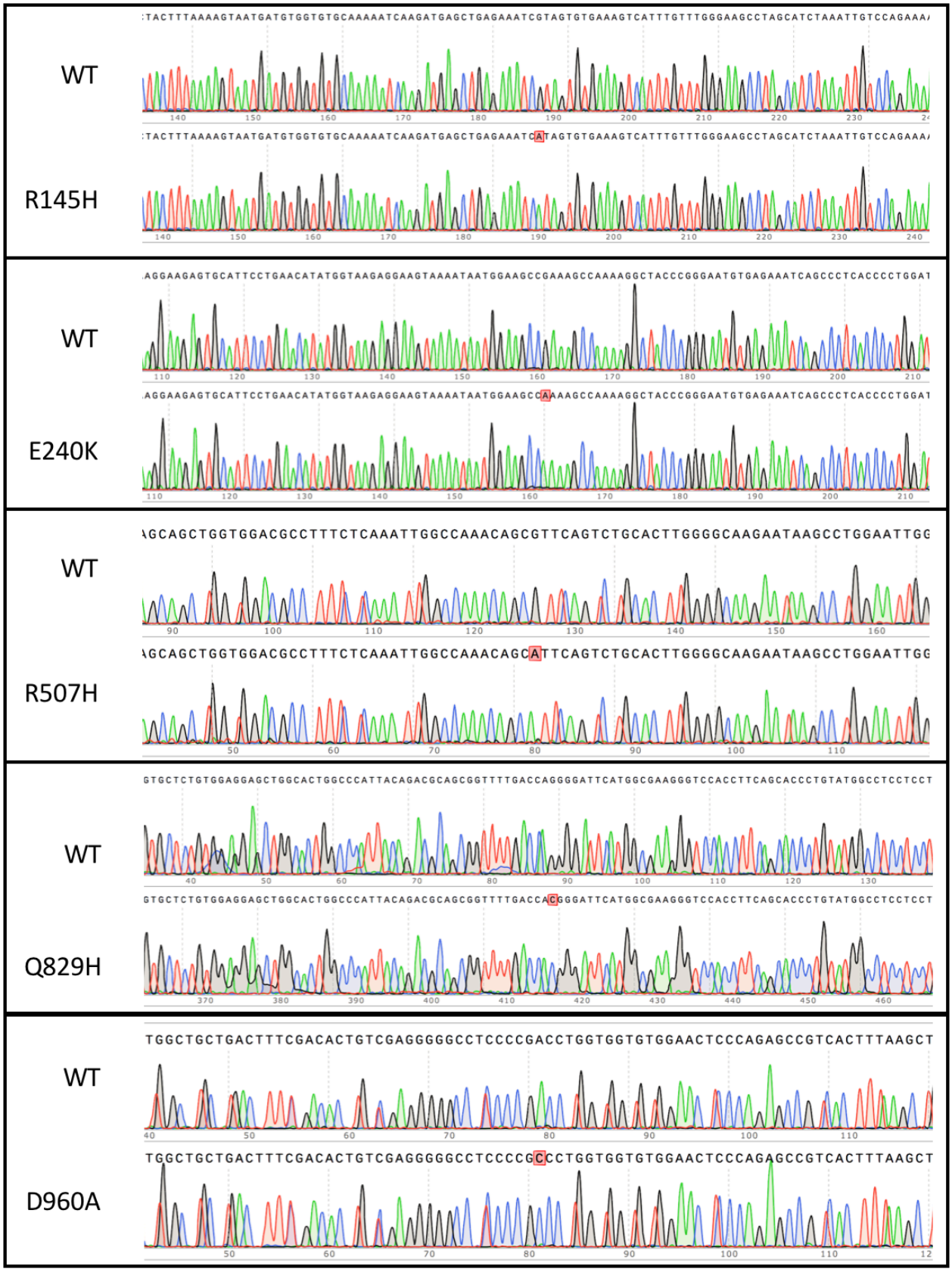


Figure 6.8. Sanger sequencing confirming SDM of the pcDNA5/FRT/TO FAN1 vector.
From top to bottom, p.R145H, p.E240K, p.R507H, p.Q829H and p.D960A are presented along with the wild type (WT) vector sequencing at each locus. Variants are marked in red boxes.

6.5.1.3.2 Titration of FAN1 expression

U20S cells express GFP-FAN1 at approximately physiological levels in the absence of induction, whereas tetracycline at the 1 µg/ml suggested by Munoz et al. (2014), induced massive overexpression. Expression at such a high level appears toxic to cells, sensitising them to MMC-induced ICLs to the same degree as *FAN1* knockout.

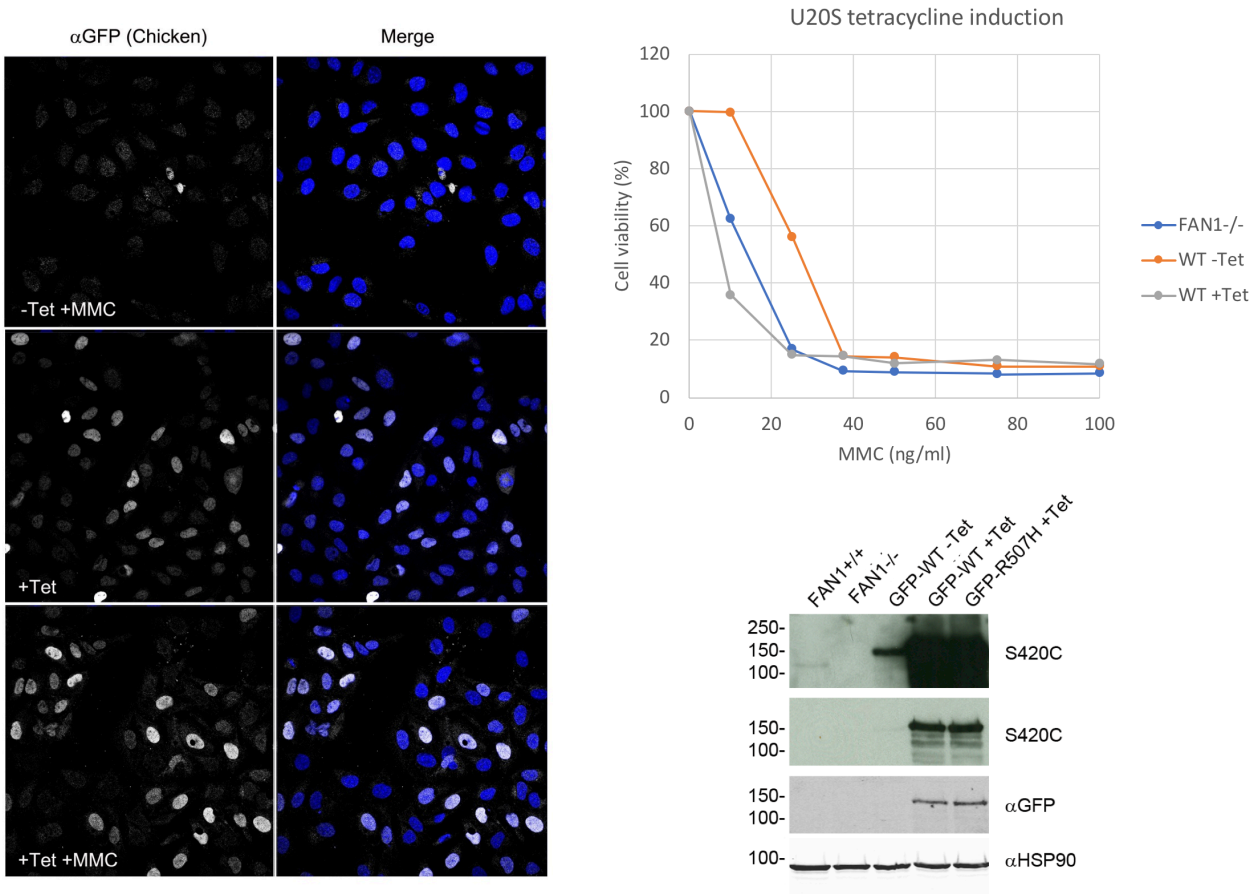


Figure 6.9. Tetracycline induction reverses the protective effect of GFP-FAN1.
Left *– confocal immunofluorescence of U20S cells transfected with wild type GFP-FAN1. Anti-GFP antibody (white), nuclei are stained with DAPI (blue). Cells have been treated with MMC (top panel), tetracycline (middle panel) or both (bottom panel).* ***Top right*** *– viability of U20S cells expressing wild type (WT) GFP-FAN1 following exposure to MMC at the indicated doses. FAN1 knockout (blue) sensitises cells to MMC, an effect reversed by complementation with wild type GFP-FAN1 (red). Tetracycline at 1 µg/ml induces overexpression of wild type GFP-FAN1 (grey), which reduced viability to knockout levels.* ***Bottom right*** *– immunoblot of tetracycline-induced GFP-FAN1 expression. Control U20S cells (FAN1+/+) and FAN1 knockout (-/-) are shown. GFP-WT is expressed at approximately physiological levels in the absence of tetracycline, but tetracycline induces overexpression of both p.R507H and wild type (WT) GFP-FAN1. Blots are probed with S420C anti-FAN1 antibody and anti-GFP antibody. Heat shock protein 90 (HSP90) is used as a loading control.*

A dose titration assessed toxicity at lower doses, but even tetracycline concentrations as low as 0.06 µg/ml had a detrimental effect on viability following MMC exposure.

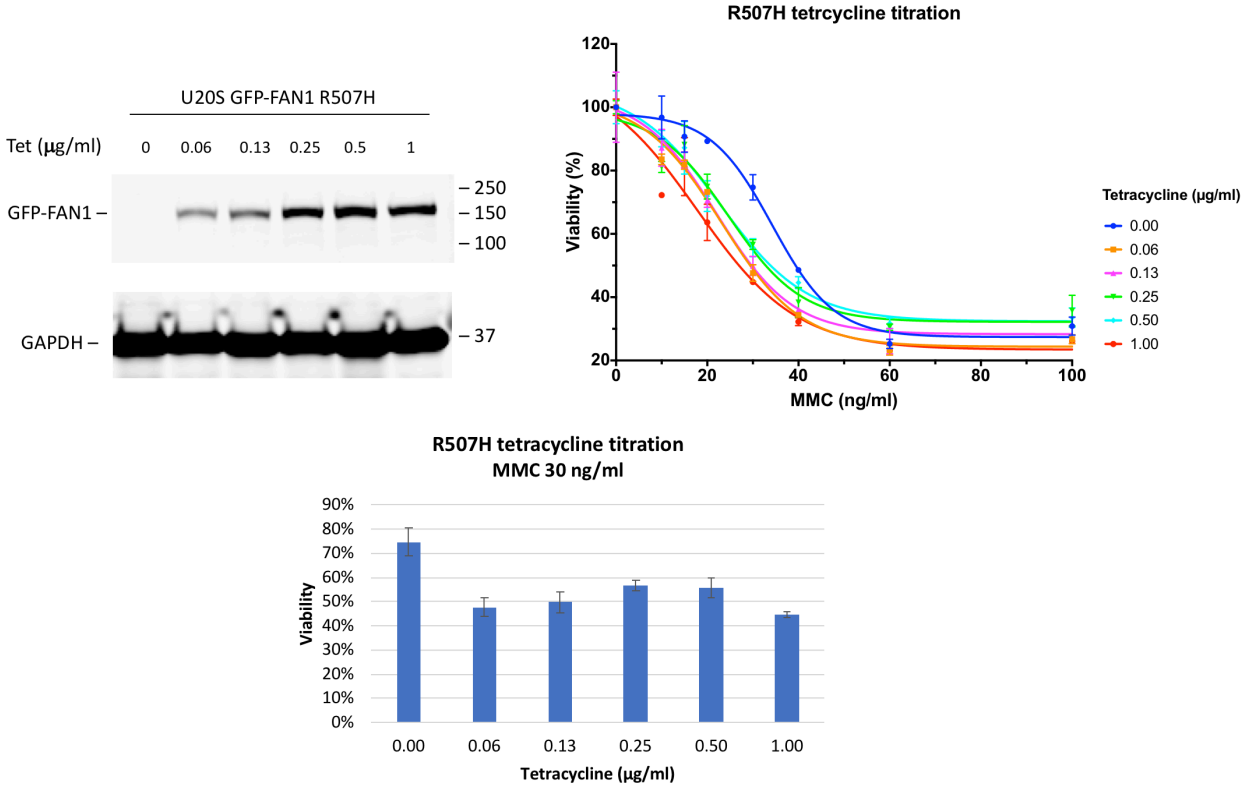


Figure 6.10. Tetracycline dose titration in p.R507H FAN1 U20S cells exposed to MMC.
Top left *–immunoblot probed with the indicated antibodies, shows GFP-FAN1 expression level increases with tetracycline dose.* ***Top right*** *– viability of U20S cells induced with the indicated dose of tetracycline and exposed to the increasing concentrations of MMC.* ***Bottom*** *– viability of U20S cells exposed to 30 ng/ml MMC and induced with increasing concentrations of tetracycline.*

6.5.1.3.3 Nuclear interstrand crosslink repair foci

Cisplatin induces ICLs, repair of which involves the induction of double strand breaks that can be visualised as nuclear foci of phosphorylated histone H2AX (γ-H2AX) (Niedernhofer et al., 2004, Rothfuss and Grompe, 2004). Foci normally peak at 24 h then clear by 48 h (MacKay et al., 2010b). In U20S cells, *FAN1* knockout did not affect the appearance of foci but did prevent their clearance at 72h. Reconstituting cells with either wild type or p.R507H GFP-FAN1 restored γ-H2AX clearance.

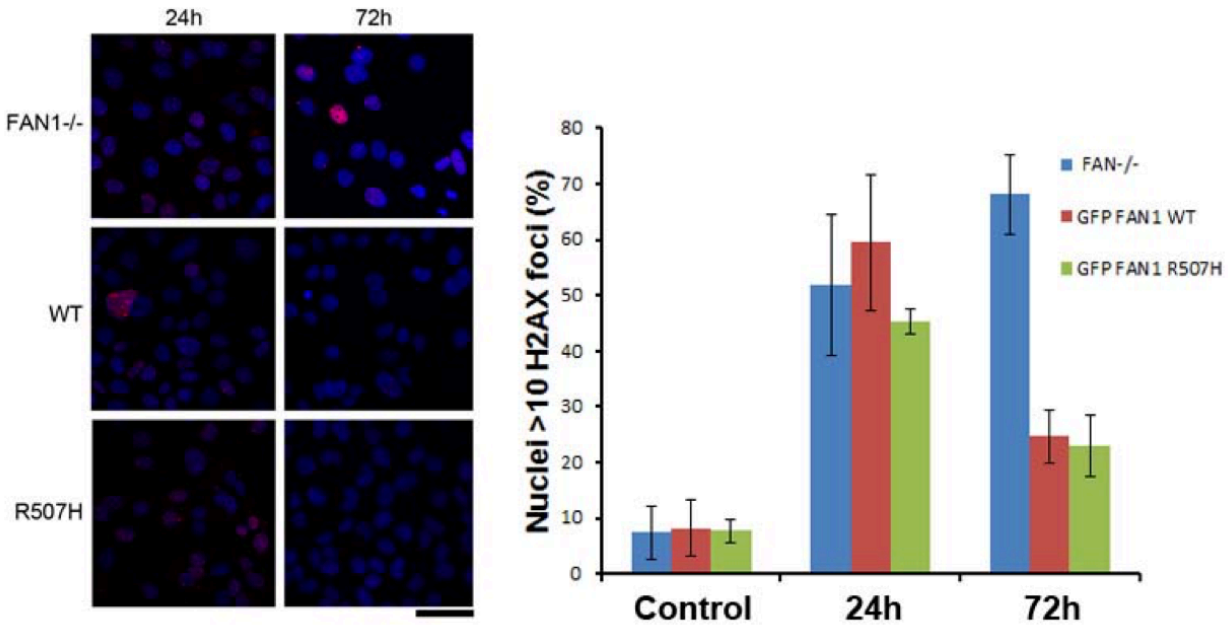


Figure 6.11. γ-H2AX assay following cisplatin exposure in U20S cells.
Left *– confocal images of FAN1 knockout (-/-) U20S cells and those complemented with wild type (WT) or R507H GFP-FAN1. Immunofluorescence for γ-H2AX foci (red) and counterstaining with DAPI nuclear stain (blue). Scale bar 50 μM.* ***Right*** *– quantification of nuclei showing >10 γ-H2AX foci (results from two separate experiments).*

6.5.1.3.4 Interstrand crosslink sensitivity

FAN1 knockout sensitises U20S cells to MMC-induced ICLs, and expression of wild type, p.R145H, p.E240K, p.R507H or p.Q829H GFP-FAN1 constructs restored resistance. As expected, the nuclease inactivated p.D960A mutation (Kratz et al., 2010a, Liu et al., 2010b), either on a WT or p.R507H *FAN1* background, compromised ICL repair function.

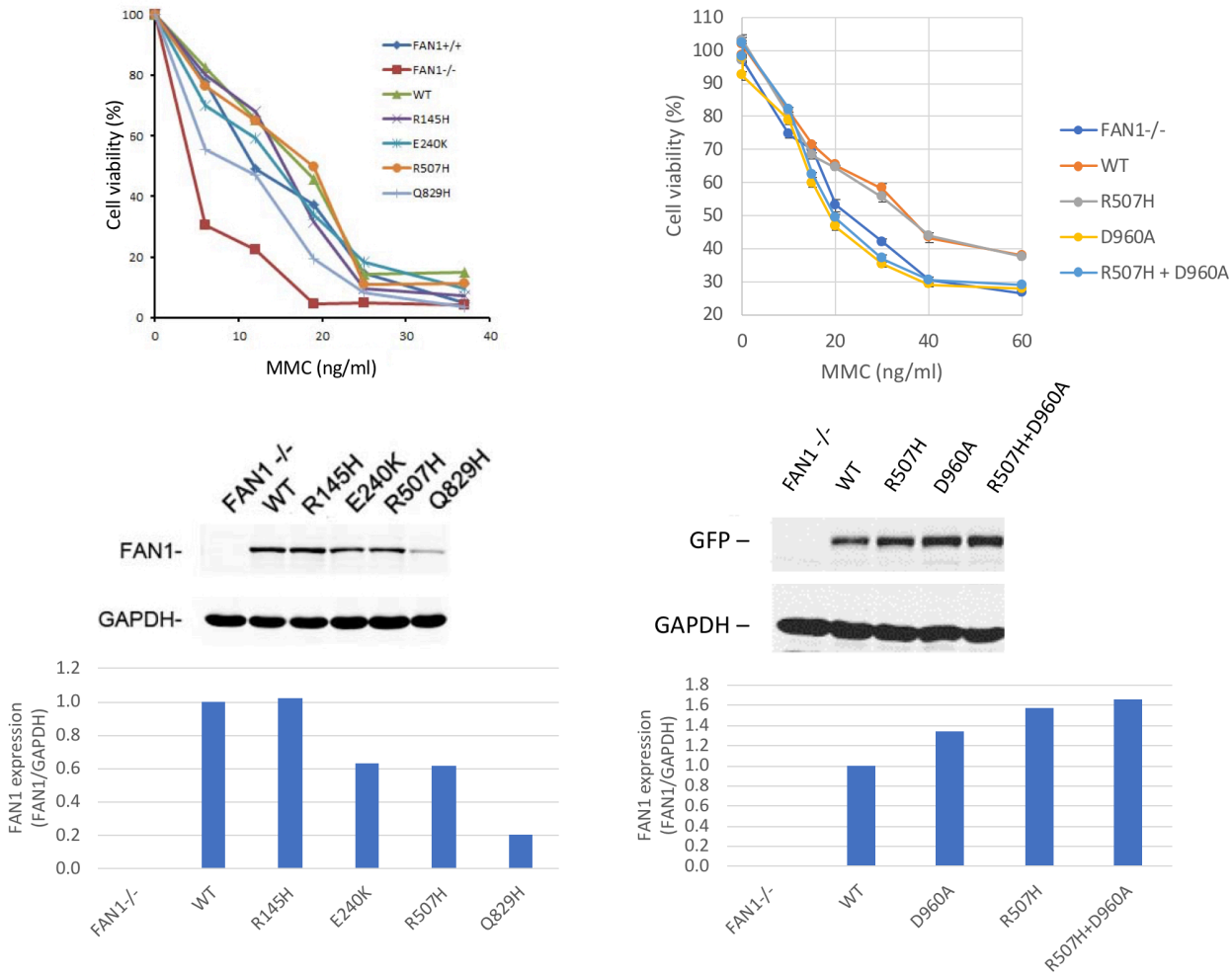


Figure 6.12. FAN1 knockout sensitises U20S cells to MMC-induced interstrand crosslinks (ICL) and expression of wild type or variant GFP-FAN1 restores resistance.

Top left *– viability assay in U20S cells exposed to MMC for 20 h. Cells express endogenous FAN1 (+/+), wild type or variant FAN1, or have FAN1 knocked out (-/-). Viability was determined by MTT assay after 10 days.* ***Bottom left*** *– immunoblot of U20S knockout (-/-) cells and those complemented with FAN1 constructs in the absence of tetracycline induction. Blots are probed with the S420C FAN1 antibody and GAPDH is used as a loading control. Quantification is corrected for GAPDH loading control and given relative to WT.* ***Top right*** *– viability assay in U20S cells with the variants indicated, exposed to MMC for 20 h. The p.D960A nuclease mutant sensitises cells to MMC.* ***Bottom right*** *– immunoblot showing FAN1 levels in U20S cells reconstituted with catalytically active or nuclease dead WT or p.R507H variant FAN1 forms. The FAN1$^{-/-}$ line is shown for comparison. Quantification is corrected for GAPDH loading control and given relative to WT. Note the p.R507H forms are expressed at a higher level.*

6.5.1.3.5 Genotoxins

Similar results were found for cisplatin, which also induces ICLs. However, *FAN1* knockout and p.R507H expression did not influence sensitivity to 6-TG, which invokes the MMR pathway, or hydrogen peroxide, which causes single and double strand breaks.

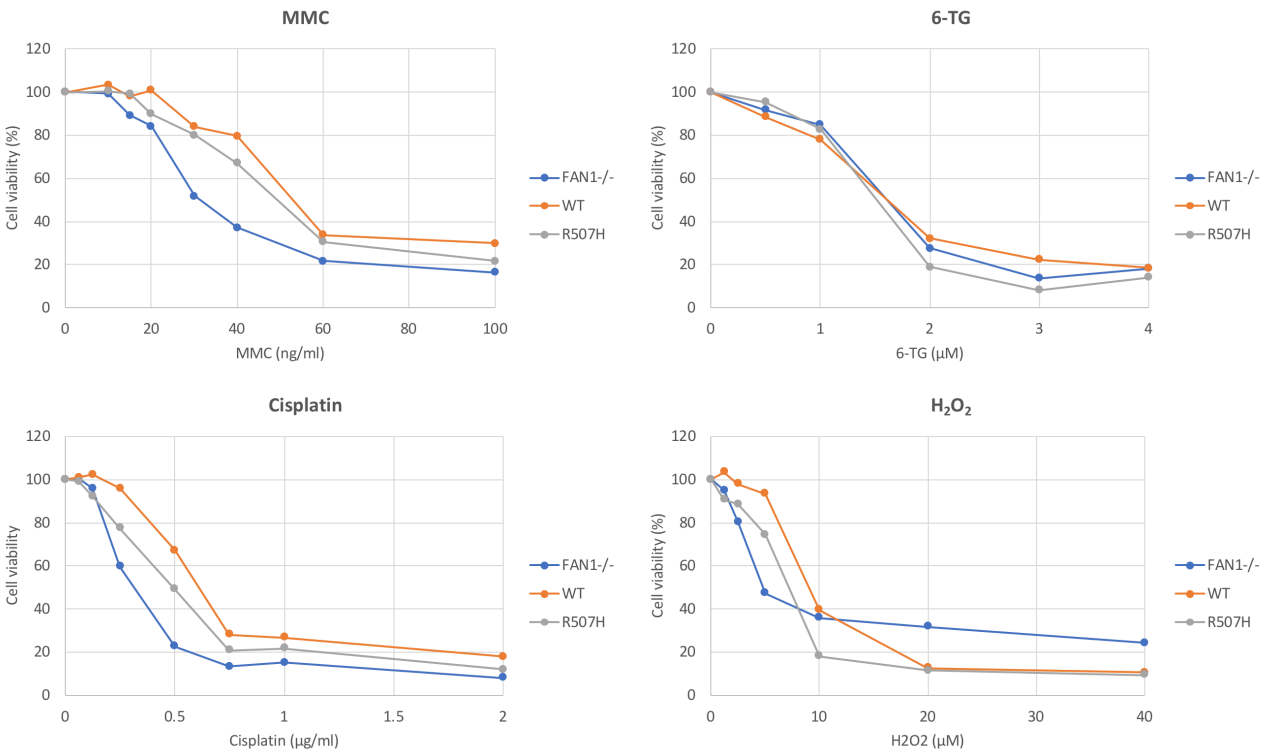


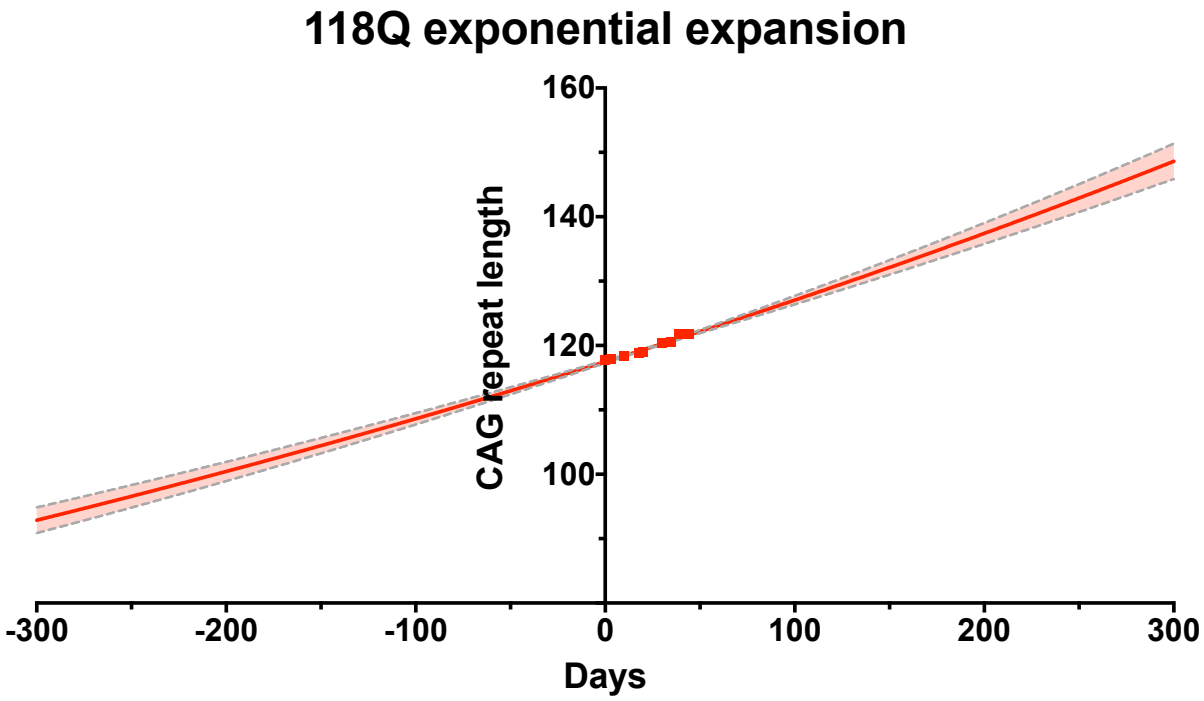
Figure 6.13. Genotoxin assays in U20S cells.

Cell viability determined by MTT assay. ***Top left*** *– U20S cells exposed to MMC at the indicated concentration for 24 h, cell viability assay 10 d later.* ***Bottom left*** *– U20S cells exposed to cisplatin at the indicated concentration for 24 h, cell viability assay 10 d later.* ***Top right*** *– U20S cells exposed to 6-TG at the indicated concentration for 24 h, cell viability assay 10 d later.* ***Bottom right*** *– U20S cells exposed to H_2O_2 at the indicated concentration for 30 min, cell viability assay 10 d later.*

6.5.2 CAG repeat instability

6.5.2.1 FAN1 expression level

The 118Q repeat in U20S FAN1$^{-/-}$ cells expanded exponentially in culture (r^2 = 0.9264, p = 8.892e-27).



ln(CAG) = (7.818e-04 * d) + 4.766

Figure 6.14. Model of U20S FAN1$^{-/-}$ 118Q exponential expansion.
r^2 = 0.9264, p = 8.892e-27. The exponential model is represented by a solid line, and the 95% confidence interval by dotted lines.

A linear model fit almost as well (r^2 = 0.9248, p = 1.435E-26), giving an expansion rate of 10.69 ± 0.44 days/Q. The CAG repeat was stabilised in cells expressing endogenous *FAN1* (FAN1$^{+/+}$) or complemented with wild type GFP-FAN1, with the expansion rate slowed to 17.70 ± 1.36 days/Q (p = 2.69E-04) and 14.83 ± 1.10 days/Q respectively (p = 2.57E-04). Overexpression of *FAN1* by induction with 0.1 µg/ml tetracycline further slowed expansion to 21.28 ± 1.84 days/Q (p = 9.86E-05). These results show increasing *FAN1* expression slows CAG expansion rate.

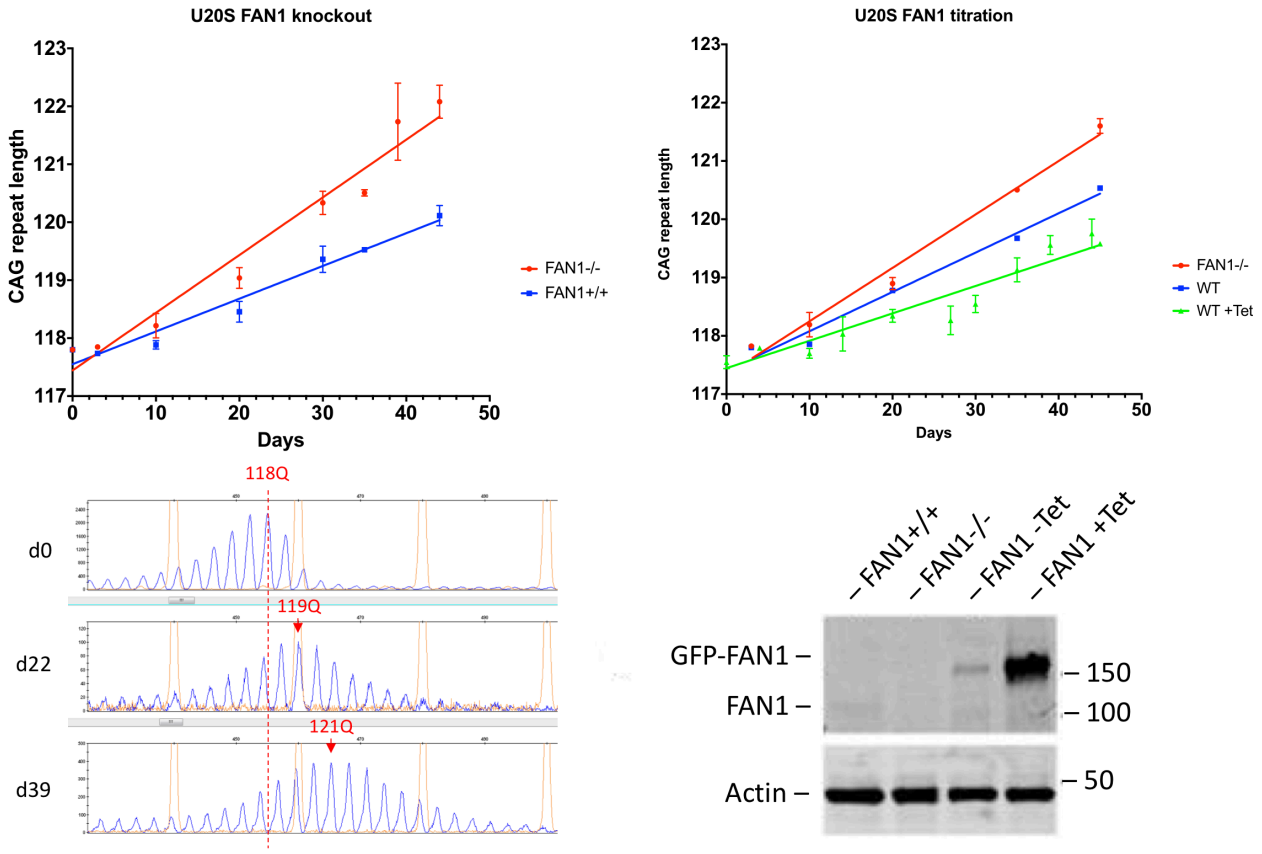
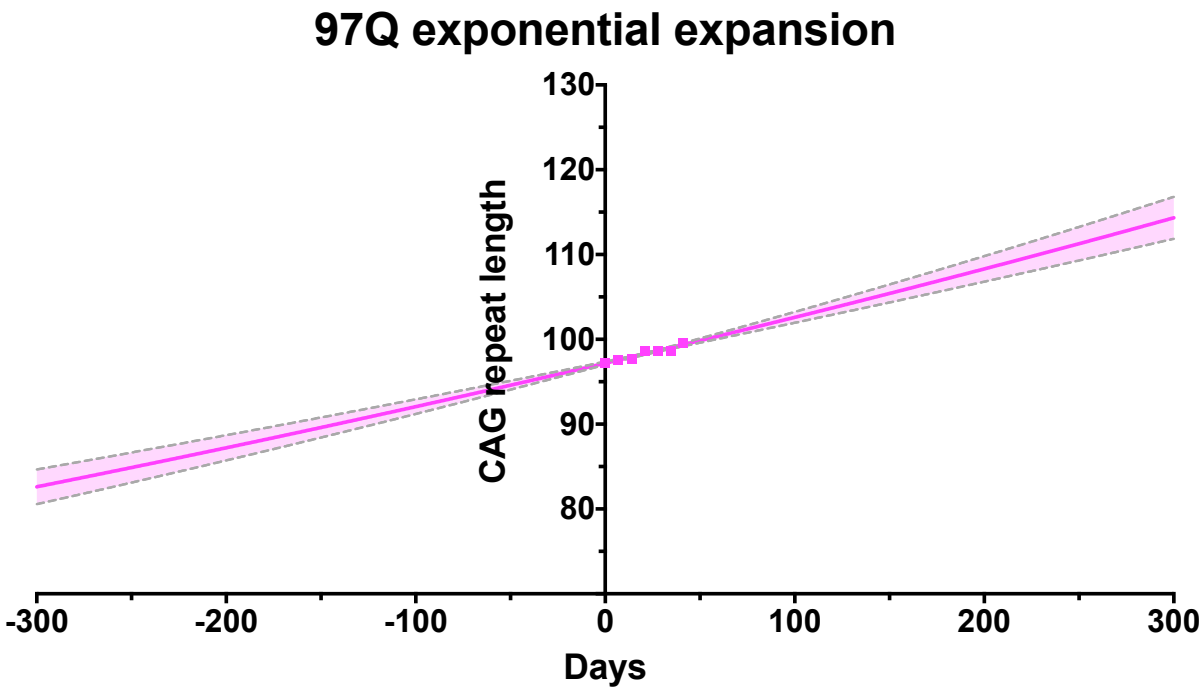


Figure 6.15. FAN1 protects against U20S 118Q CAG repeat expansion in a dose-dependent manner.
Top left *– modal CAG repeat length in FAN1$^{-/-}$ and FAN1$^{+/+}$ cells stably transduced with 118Q HTT exon 1. Data from 6x FAN1$^{-/-}$ and 5x FAN1$^{+/+}$ cultures are combined.* ***Top right*** *– modal CAG repeat length in FAN1$^{-/-}$ or FAN1$^{GFP\text{-}WT}$ cells (± 0.1 µg/ml tetracycline induction) stably transduced with 118Q HTT exon 1. Data from 5x FAN1$^{-/-}$ and 5x FAN1^{WT} and 11x FAN1^{WT} +Tet cultures are combined. Error bars represent SEM.* ***Bottom left*** *– representative repeat sizing in U20S FAN1$^{-/-}$ cells stably expressing 118Q HTT exon 1, samples are taken on the days indicated following transduction.* ***Bottom right*** *– Cell lysates for SDS-PAGE from U20S FAN1$^{+/+}$, FAN1$^{-/-}$, and FAN1$^{GFP\text{-}WT}$, with or without induction of expression by 0.1 µg/ml tetracycline (Tet). Antibodies to GFP, FAN1 (S420C) and actin are shown. In the absence of Tet, low levels of GFP-FAN1 are expressed, close to endogenous expression levels seen in FAN1$^{+/+}$ cells. Tetracycline induces FAN1 overexpression.*

6.5.2.2 CAG length dependence

30 and 70Q *HTT* exon 1 fragments were stable in U20S FAN1$^{-/-}$ cells throughout 8 weeks in culture. The modal CAG length of the 97Q construct expanded exponentially (r^2 = 0.8795, p = 1.690E-08), though again fit almost as well to a linear model (r^2 = 0.8793, p = 1.711E-08) with an expansion rate of 18.41 ± 1.55 days/Q. This was significantly slower than the 118Q construct (p = 1.38E-05). These results show CAG expansion is detectable over a relatively short time in culture, provided long repeat lengths are used, and suggests expansion rate is CAG length dependent.



ln(CAG) = (5.516e-04 * d) + 4.576

Figure 6.16. Model of U20S FAN1$^{-/-}$ 97Q exponential expansion.
r^2 = 0.8795, p = 1.690e-08. The exponential model is represented by a solid line, and the 95% confidence interval by dotted lines.

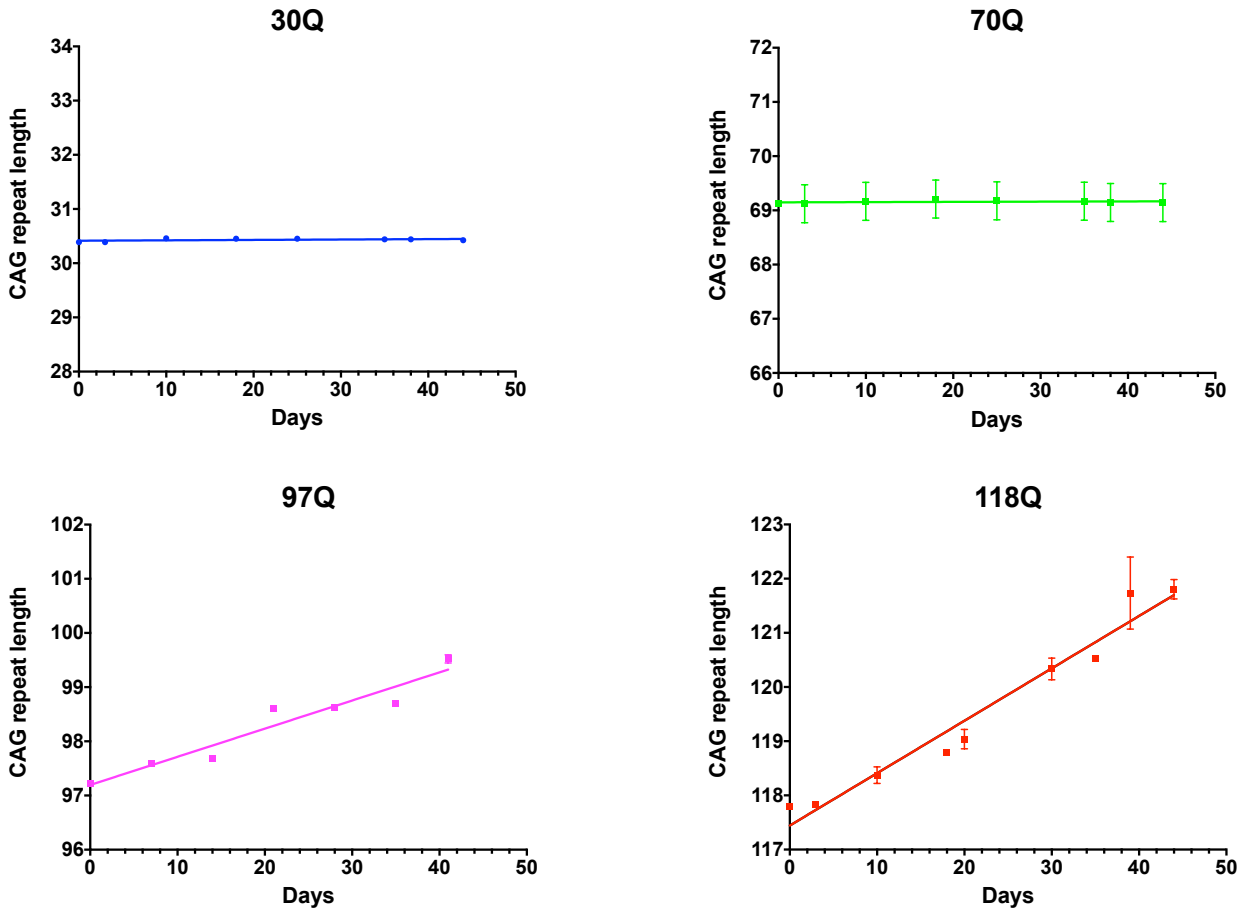


Figure 6.17. CAG repeat expansion in U20S FAN1$^{-/-}$ cells is length dependent.
Panels show modal CAG repeat length in FAN1$^{-/-}$ cells stably transduced with HTT exon 1 containing the indicated CAG repeat length. 30 and 70Q are stable over the 40-day culture. 118Q expands, as shown above. 97Q also expands, though at a slower rate than 118Q. Data from 3 cultures for each repeat length are combined, error bars represent SEM.

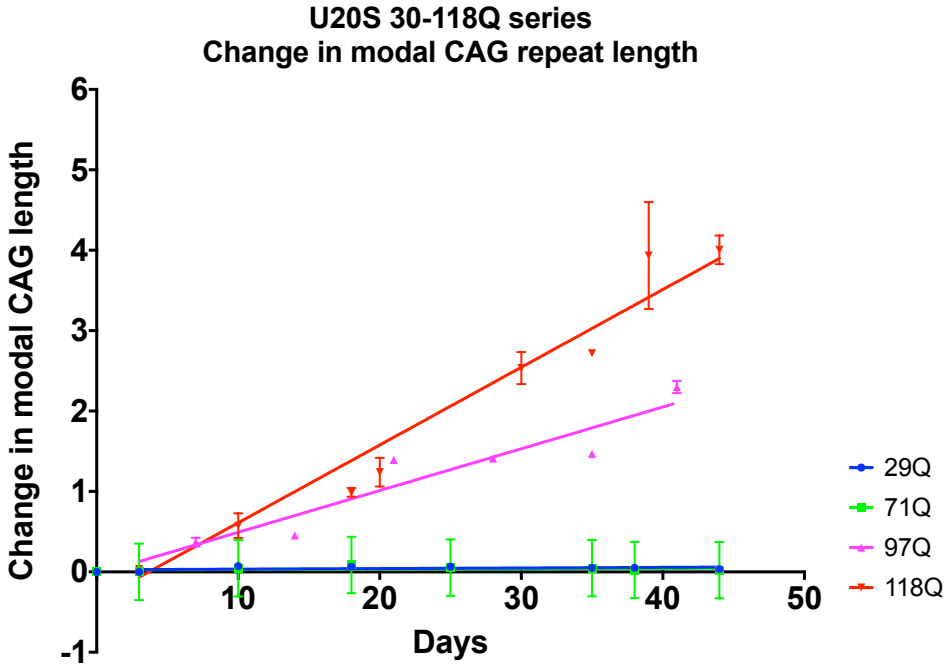


Figure 6.18. Change in modal CAG repeat length of U20S FAN1-/- cells expressing HTT exon 1 with 29-118Q.

The 118Q exponential model (p = 8.892E-27) predicts the 30Q construct would expand at 41.38 days/Q and the 70Q construct at 18.37 days/Q in the absence of FAN1. The 97Q exponential model (p = 1.690E-08) predicts they would expand at 58.64 and 26.04 days/Q respectively. Both 30Q and 70Q constructs were cultured for 8 weeks without identifiable CAG repeat length change, suggesting further data over a longer time period is required to improve the accuracy of exponential models.

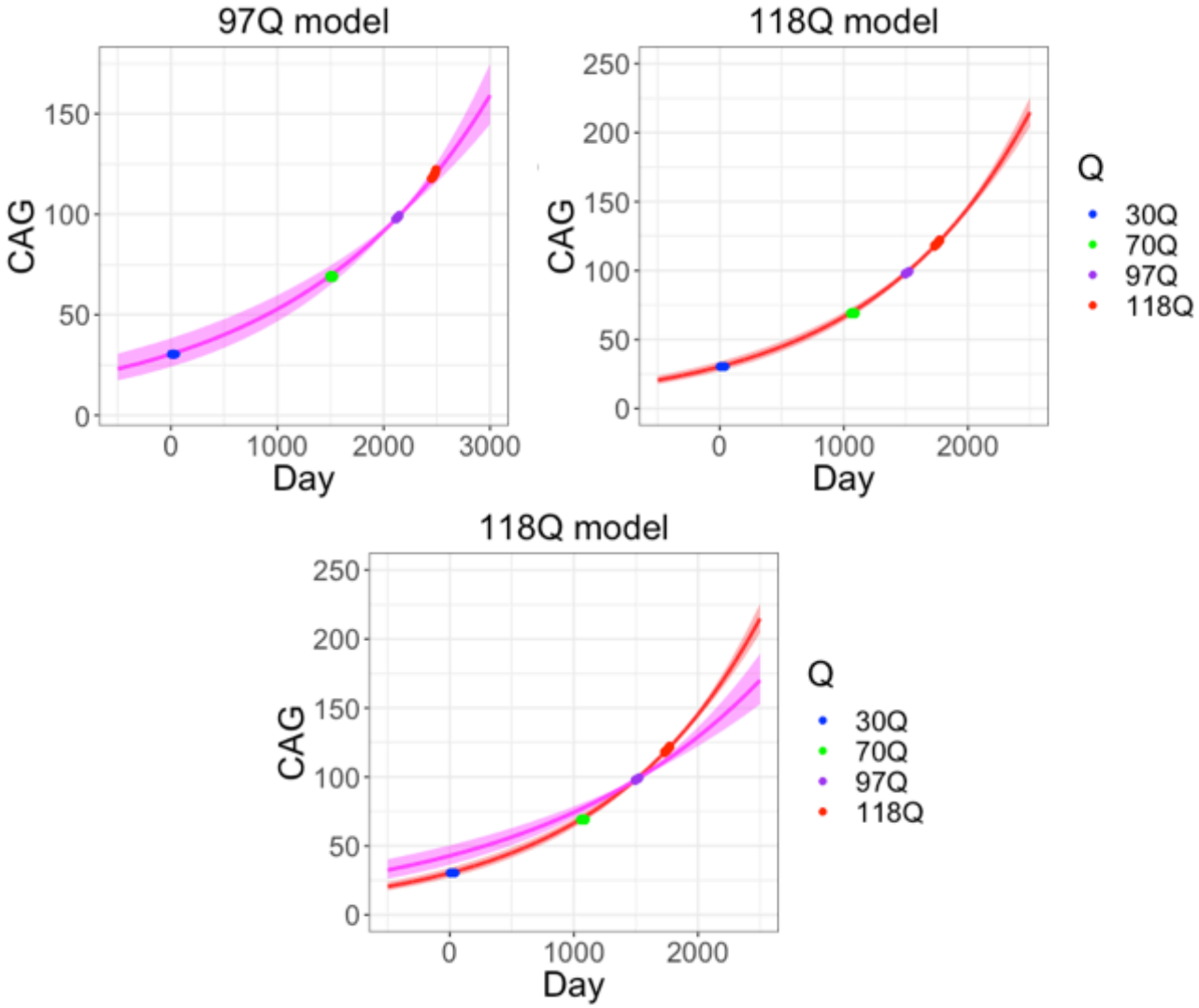


Figure 6.19. Overlay of U20S FAN1$^{-/-}$ exponential expansion.
The indicated exponential model was used to offset culture start date for each cell line. Exponential model predictions are indicated by a solid line and the 95% confidence interval by the shaded area. Orange – 30Q, green – 70Q, blue – 97Q, purple – 118Q. ***Top left*** *– exponential model of U20S FAN1$^{-/-}$ 97Q; ln(CAG) = (5.516e-04 * d) + 4.576; r^2 = 0.8795, p = 1.690e-08.* ***Top right*** *– exponential model of U20S FAN1$^{-/-}$ 118Q; ln(CAG) = (7.818e-04 * d) + 4.766; r^2 = 0.9264, p = 8.892e-27.* ***Bottom*** *–118Q exponential model with the 97Q model overlain.*

The p.R507H DNA binding domain variant did not significantly alter CAG expansion rate relative to FAN1^{WT} (p = 0.0878). The p.D960A nuclease dead mutant, either alone or with p.R507H, did not affect the rate of expansion relative to FAN1^{WT} or FAN1^{R507H} (p = 0.331 and p = 0.882 respectively), suggesting the nuclease domain is not required for FAN1 to stabilise the *HTT* CAG repeat.

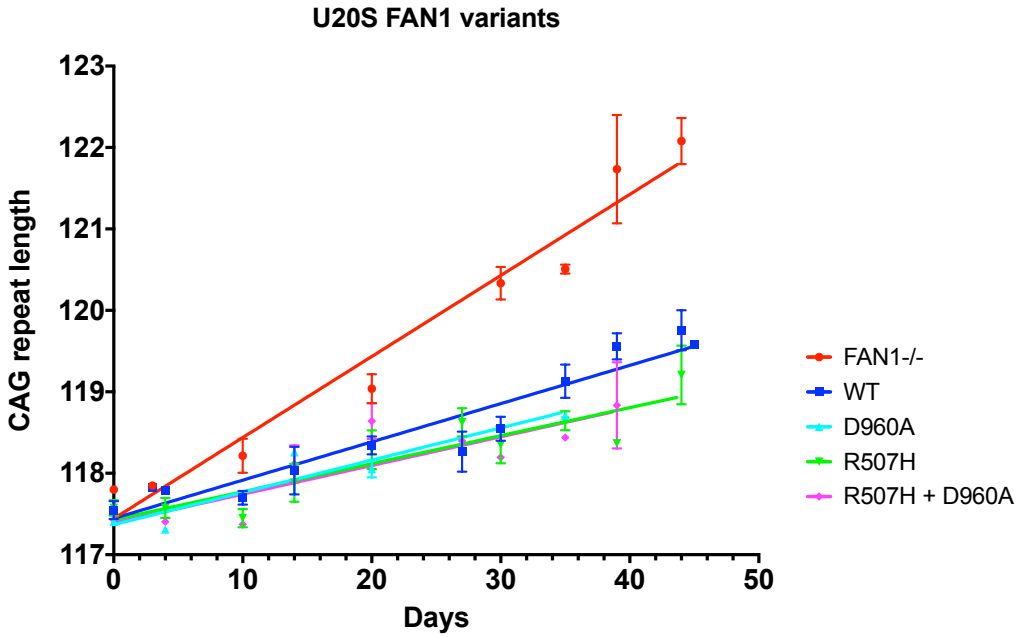


Figure 6.20. FAN1 nuclease and p.R507H variants do not modify CAG repeat expansion rate.
Modal CAG repeat size in FAN1$^{-/-}$ cells reconstituted with the FAN1 forms as indicated and stably transduced with 118Q HTT exon 1. Linear trend lines have been fitted. Data combined from 6x FAN1$^{-/-}$, 12x WT, 3x D960A (nuclease inactivated), 7x R507H and 4x R507H+D960A cultures, error bars represent SEM. Expression levels of each variant line are shown above, note p.R507H forms are expressed at a higher level.

6.5.3 FAN1-DNA interaction

To investigate for an interaction between FAN1 and *HTT* DNA, chromatin fractions were prepared from HD cell lines. Endogenous FAN1 was immunoprecipitated using the S420C antibody and the presence of *HTT* CAG DNA was assayed by qPCR.

6.5.3.1 U20S cells

U20S FAN1$^{-/-}$ and cells expressing wild type *FAN1* were transiently transfected with 70Q *HTT* exon 1, aiming to introduce high copy numbers of plasmid DNA. *FAN1* expression was induced by tetracycline in order to optimise conditions for detecting an interaction. PCR detected low, but reproducible levels of CAG repeat DNA in ChIP fractions and amplified both the normal and expanded *HTT* CAG alleles. qPCR primers spanning the *HTT* CAG repeat demonstrated an enrichment in U20S FAN1^{WT} ChIP fractions relative to the control immunoprecipitation (IP) without antibody, and as expected there was no significant enrichment in FAN1$^{-/-}$ cells. This suggests a novel interaction between FAN1 and CAG repeat DNA, and demonstrates the specificity of the ChIP pull down. However, primers targeting *HTT* DNA near (HTT con 1, intron 3-4) and far from the CAG repeat (HTT con 2, exon 49) also showed enrichment in ChIP fractions, suggesting FAN1 binds DNA but does not specifically target the CAG repeat.

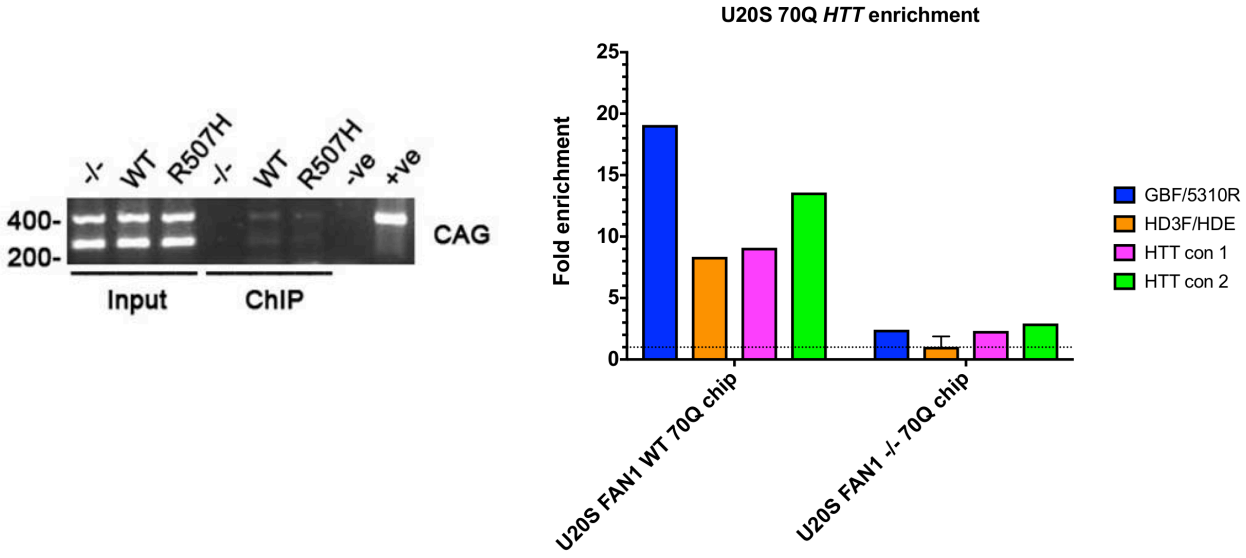


Figure 6.21. FAN1 binds HTT CAG repeat DNA.

U20S FAN1$^{-/-}$ cells and those reconstituted with WT FAN1 were transiently transfected with 70Q HTT exon 1. ChIP with the sheep S420C antibody was used to isolate FAN1. ***Left*** *– agarose gel of input and ChIP fractions amplified with HD3F/HD5 primers which span the CAG repeat. Negative (water) and positive (70Q HTT exon 1 plasmid) are shown. The smaller band represents the endogenous HTT allele and the larger is 70Q exon 1.* ***Right*** *– SYBR green qPCR of ChIP fractions using the primers indicated. Enrichment is given relative to the control IP without antibody (dotted line). GBF/5310R and HD3F/HDE span the CAG, HTT con 1 primers are 25 kb 3' of the CAG repeat, between exons 3 and 4, and HTT con 2 primers are 138 kb 3 of the repeat', at exon 49. All samples run in triplicate, error bars represent SEM of independent runs.*

6.5.3.2 Patient-derived induced pluripotent stem cells (iPSC)

ChIP fractions were prepared from 109Q and 125Q iPSCs, which are homozygous for wild type *FAN1* and contain unstable *HTT* CAG repeats. Once again PCR amplified both the normal and pathogenic CAG repeat. qPCR showed significant enrichment of *HTT* CAG repeat DNA (p 109Q = 0.0208, p 125Q = 1.672E-03), but also enriched for distal regions of *HTT* (p 109Q = 3.068E-03) and the CAG repeat DNA of *ATXN3* and *TBP*.

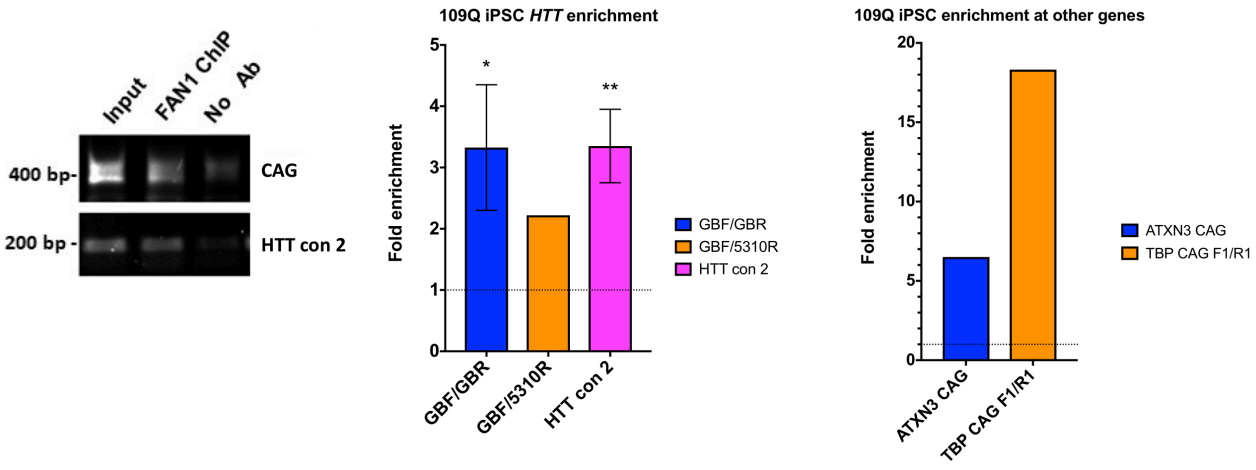


Figure 6.22. FAN1 interacts with endogenous HTT DNA of 109Q iPSCs.
Left *– 109Q iPSC input, FAN1 ChIP and no antibody control fractions were PCR amplified with a primer pair spanning the CAG repeat (HD3F/HD5) and a pair at the 3' end of the HTT gene (HTT con 2), then run on an agarose gel. Input (5%) and ChIP fractions are shown. The two bands seen in CAG PCR are from the normal and pathogenic alleles.* ***Middle*** *– SYBR green qPCR with primer pairs spanning the CAG repeat (GBF/GBR and GBF/5310R) and a pair at the 3' end of the HTT gene (HTT con 2). Enrichment is given relative to the control IP without antibody (dotted line).* ***Right*** *– qPCR with primer pairs spanning the CAG repeats in ATXN3 and TBP. All samples run in triplicate, two independent experiments for HTT qPCR, error bars represent SEM.*

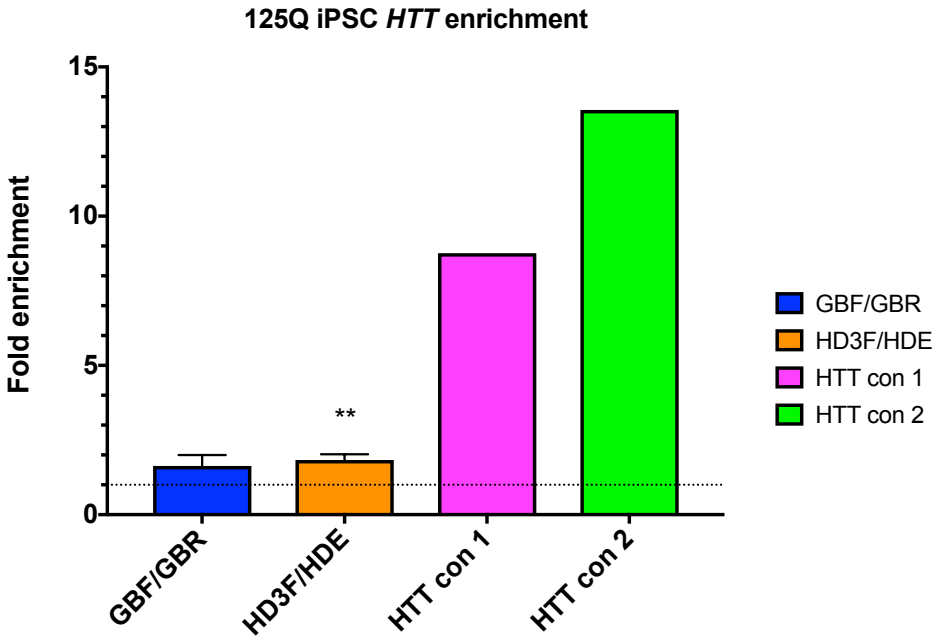


Figure 6.23. FAN1 interacts with endogenous HTT DNA of 125Q iPSCs.
SYBR green qPCR with primer pairs spanning the CAG repeat (GBF/GBR and GBF/5310R), and pairs distal to the repeat; HTT con 1 primers bind between exons 3 and 4, and HTT con 2 primers bind at exon 49. Enrichment is given relative to the control IP without antibody (dotted line). All samples run in triplicate, 2-4 independent experiments for CAG repeat qPCR, error bars represent SEM.

6.5.3.3 Patient-derived lymphoblastoid cells

FAN1 ChIP fractions were prepared from two lymphoblastoid (LB) cell lines either homozygous for wild type *FAN1* or heterozygous for the p.R507H variant. PCR with primers spanning the CAG repeat showed amplification of both the short and long alleles. qPCR with primers spanning the *HTT* CAG repeat showed enrichment in ChIP fractions from both wild type and p.R507H lines (p WT = 0.0158, p R507H = 0.0316). There was also significant enrichment for two *HTT* primer

pairs distal to the CAG repeat (p HTT con 1 = 0.0155, p HTT con 2 = 0.0129) and for primers targeting trinucleotide repeat and control regions of *ATXN3*, *DMPK*, *FXN* and *TBP* genes. There was no significant difference in enrichment at any of the targets between wild type and p.R507H lines.

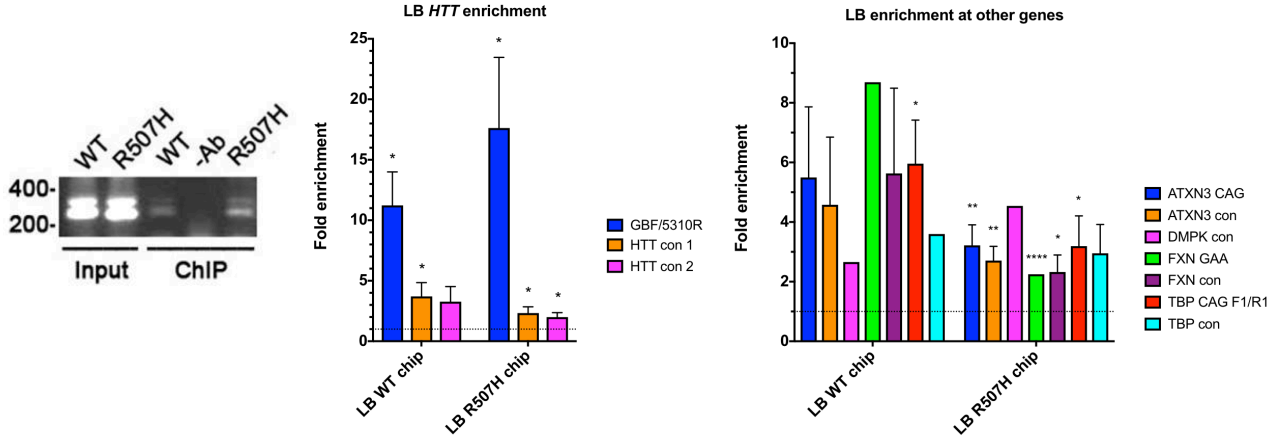


Figure 6.24. FAN1 interacts with endogenous HTT DNA of HD lymphoblasts (LB).
Left *– input, FAN1 ChIP and no antibody control fractions from HD lymphoblasts derived from patients homozygous for wild type FAN1 or heterozygous for the p.R507H variant were PCR amplified by primers spanning the CAG repeat (HD3F/HD5), then run on an agarose gel. Input (5%) and ChIP fractions are shown. The two bands seen in CAG PCR are from the normal and pathogenic alleles.* ***Middle*** *– SYBR green qPCR with primers spanning the CAG repeat (GBF/5310R) and pairs distal to the repeat; HTT con 1 primers are between exons 3 and 4, and HTT con 2 primers are at exon 49. Enrichment is given relative to the control IP without antibody (dotted line).* ***Right*** *– qPCR with primer pairs spanning trinucleotide repeat and control regions of ATXN3, DMPK, FXN and TBP. Samples in triplicate, each run 2-7 times independently, error bars represent SEM.*

6.6 Discussion

6.6.1 Interstrand crosslink repair

As expected, siRNA-mediated FAN1 knockdown in HEK 293 cells increased sensitivity to MMC-induced interstrand crosslinks (ICL). Stably expressed Myc-tagged full length wild type or p.R507H FAN1 formed nuclear DNA repair foci after ICL induction, demonstrating that the artificial constructs are functional in ICL repair, but that the DNA binding domain variant, which is associated with early HD onset, does not affect FAN1 localisation to ICL lesions. A HD patient-derived lymphoblastoid line heterozygous for the p.R507H variant appeared resistant to MMC-induced ICLs, relative to most control lines, though this may have been due to its increased *FAN1* expression level. *FAN1* knockout in U20S cells sensitised them to ICLs and delayed resolution of double strand breaks (DSB) formed during ICL repair, but did not affect mismatch repair function. Complementation with p.R145H, p.E240K, p.R507H or p.Q829H FAN1 restored resistance to wild type levels, again demonstrating full functionality of the variant lines in ICL repair.

6.6.2 Repeat instability

6.6.2.1 FAN1 stabilises the HTT CAG repeat

U20S cells stably transduced with 118Q *HTT* exon 1 show repeat expansion that is significantly accelerated by *FAN1* knockout ($FAN1^{-/-}$, p = 2.57E-04). Increasing *FAN1* expression incrementally slowed CAG expansion rate, suggesting its stabilising effect is dose-dependent. This is consistent with a recent study that showed *Fan1* knockout accelerates CGG repeat expansion in a mouse model of fragile X syndrome. The p.R507H variant did not significantly alter expansion rate relative to wild type *FAN1* in U20S cells.

FAN1 is likely to play a key role in a network of DNA damage response (DDR) proteins, as suggested by GWAS pathway analyses finding significant association of DNA repair gene sets with onset (GeM-HD, 2015). These include mismatch repair components *MSH3* and *MLH1*, inactivation of which in HD mouse models abrogates somatic expansion and ameliorates the HD phenotype (Pinto et al., 2013a, Tome et al., 2013a). The *in vitro* cell models presented here suggest *FAN1* also contributes to this mechanism, protecting against CAG expansion.

6.6.2.2 CAG length dependence

A non-pathogenic 30Q and a pathogenic 70Q CAG repeat were stable in the U20S $FAN1^{-/-}$ system, which is consistent with observations in iPSCs derived from an HD patient with 73Q, which were stable in long term culture (chapter 5). A 97Q repeat expanded exponentially in culture, though at a slower rate than the 118Q repeat (p = 1.38E-05). This suggests that, in this system and over the short 6 week culture, CAG repeat expansion is repeat length-dependent, with a trigger between 70 and 97 CAG repeats.

Repeat instability likely involves the formation of unusual DNA structures such as hairpin loops (Gacy et al., 1995), with high complementarity of the repeating sequence leading to slippage or mis-hybridisation (Pearson et al., 2005b). These structures are recognised by DNA repair proteins, particularly the MMR pathway, which may trigger attempts at repair that lead to the incorporation of additional CAG units. Longer CAG repeats are more unstable and prone to increased rates of expansion (Pearson et al., 2005b), features which are recapitulated in the U20S CAG repeat series and 109Q iPSCs (Chapter 5). Cells with more typical pathogenic CAG repeat lengths, up to 73Q, remained stable in culture, which may reflect the rarity of expansion events in shorter repeats, and it may be that expansion could be detected in longer term

culture or in cell types other than U20S. HD typically manifests in midlife, which may reflect slow expansion of the CAG repeat over decades until a toxic threshold is reached, after which time the polyglutamine tract confers toxicity in vulnerable cells, such as striatal medium spiny neurons, resulting in clinical onset (Kaplan et al., 2007, Kennedy et al., 2003, Swami et al., 2009).

6.6.2.3 Nuclease activity

The p.D960A active site mutation completely inactivates FAN1 nuclease function, as demonstrated by increased sensitivity of U20S FAN1^{D960A} cells to MMC-induced ICLs to the same degree as *FAN1* knockout. However, this mutation did not affect CAG expansion rate in U20S cells expressing 118Q *HTT* exon 1, relative to wild type FAN1. This suggests the nuclease activity is not required for FAN1 to stabilise the *HTT* CAG repeat. All FAN1 functions identified to date, including ICL repair (MacKay et al., 2010b, Kratz et al., 2010a, Liu et al., 2010b, Smogorzewska et al., 2010a, Thongthip et al., 2016) and replication fork recovery (Lachaud et al., 2016a, Chaudhury et al., 2014, Porro et al., 2017), depend on its nuclease activity, suggesting a novel mechanism underlies CAG repeat stabilisation.

6.6.3 FAN1-DNA interaction

Chromatin immunoprecipitation (ChIP) suggests FAN1 binds at or near the *HTT* CAG repeat, potentially recognising unusual DNA structures (McMurray, 2010). However, FAN1 also bound other DNA regions, suggesting it does not preferentially interact with expanded CAG repeats. These results were consistent across synthetic and patient-derived cell lines, showing FAN1 also interacts with the endogenous *HTT* CAG repeat. There was no significant difference in enrichment between wild type and p.R507H FAN1, suggesting this variant does not alter its binding to these DNA substrates.

6.6.4 FAN1 function

The mechanism by which FAN1 protects against CAG repeat expansion is unknown, but three models can be postulated. **Firstly**, FAN1 may bind CAG DNA, blocking access of other DNA repair proteins such as MSH3 and preventing error-prone repair. **Secondly**, as MutLα (MLH1/PMS2) independently binds FAN1 and MutSβ (MSH2/MSH3) (MacKay et al., 2010b), FAN1 could sequester MLH1 that would otherwise act with MSH3 to promote repeat expansion. **Finally**, FAN1 could act on the CAG repeat, promoting accurate repair either directly or as a scaffold for a repair complex. Supporting this, the FAN1 interactome includes several modifiers of CAG repeat stability, including MutL components (MLH1, MLH3, and PMS2) and PCNA (MacKay et al., 2010b, Porro et al., 2017). FAN1 is recruited to stalled replication forks by ubiquitinated PCNA (Porro et al., 2017), but no function has yet been identified for the interaction with MLH1, despite evidence indicating this complex is stable and comprises a substantial proportion of the cellular FAN1 and MLH1 under steady state conditions (MacKay et al., 2010b). FAN1 has the potential DNA binding and nuclease activity to act directly at CAG repeat DNA (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Smogorzewska et al., 2010a), but results presented in this chapter suggest FAN1 activity to stabilise CAG repeats is independent of its nuclease activity.

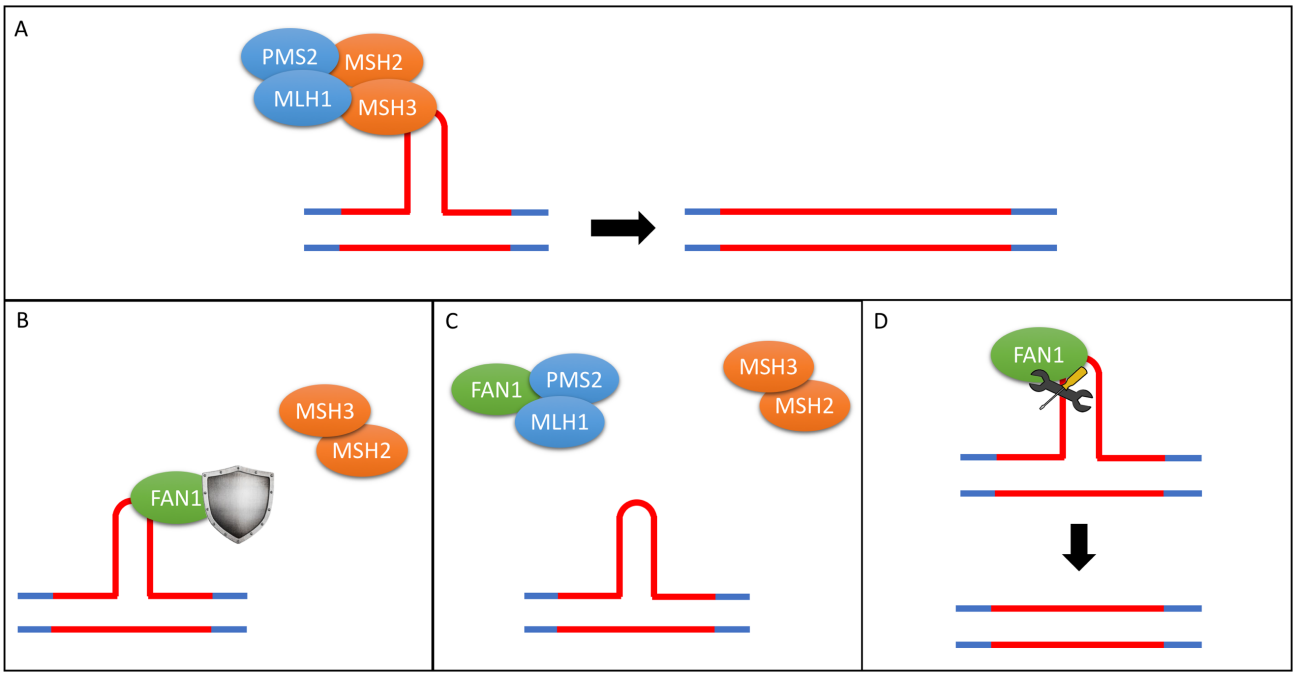


Figure 6.25. Potential mechanisms by which FAN1 may protect against CAG repeat expansion.
A) *MutSβ (MSH2/MSH3, orange) and MutLα (MLH1/PMS2, blue) misidentify abnormal secondary structures formed by CAG repeat DNA (red), such as hairpins, invoking mismatch repair, during which out of register alignment introduces repeat expansion.* ***B)*** *FAN1 (green) may bind CAG repeat DNA, prohibiting access by MutSβ.* ***C)*** *FAN1 may sequester MutLα (MLH1/PMS2) away from MutSβ, preventing MMR.* ***D)*** *FAN1 may act directly at the CAG repeat, promoting accurate repair.*

6.7 Summary

FAN1 is a nuclease involved in DNA interstrand crosslink (ICL) repair that has recently been shown to stabilise CGG repeat DNA in a mouse model of Fragile X syndrome. The results presented here validate FAN1 involvement in ICL repair and show that *FAN1* expression reduces *HTT* CAG repeat expansion. Expansion is CAG repeat length and FAN1 concentration dependent, but does not require its nuclease activity. FAN1 binds *HTT* CAG repeat DNA, but is not specifically targeted to it. These data provide new mechanistic insights into how FAN1 acts to alter disease progression. The known FAN1 interactome suggests it acts in concert with other DNA damage response (DDR) proteins, potentially sequestering MLH1 that would otherwise act with MSH3 to promote repeat expansion.

Preliminary results in cells expressing p.R507H (rs150393409), the most significant FAN1 coding SNP in the GeM GWAS of HD onset (GeM-HD, 2015), do not find any functional changes that could explain its effect on disease course. ICL repair was unaffected by its heterozygous expression in patient-derived lymphoblastoid (LB) cells or by its expression in U20S FAN$^{-/-}$ cells. CAG expansion rate in U20S$^{p.R507H}$ did not differ significantly from FAN1^{WT} cells and ChIP-qPCR showed no significant alteration in its interaction with *HTT* CAG repeat DNA. However, the semi-quantitative nature of these assays and the relatively high levels of *FAN1* expressed in U20S cells may obscure subtle differences in activity conferred by p.R507H, especially given the strong effect of expression level on expansion rate.

Further work will investigate the molecular mechanism by which FAN1 modifies CAG repeat stability, and whether this is a potential therapeutic target in HD. It will study the effect of *FAN1* variation on DNA binding and repeat expansion in U20S cells, look for differential binding of the short and long *HTT* CAG allele by tapestation or bioanalyzer analysis of ChIP-qPCR products, probe the protein interactions of FAN1 by immunoprecipitation, initially focusing on MMR

components, and will study these functions in more physiological cell models including differentiated medium spiny neurons derived from *FAN1* variant-carrying HD patients.

6.8 Publications relating to this chapter

The work presented in this chapter was published in:

FAN1 modifies Huntington's disease progression by stabilising the expanded HTT CAG repeat. Goold, R.*, **Flower, M.***, Moss, D. H., Medway, C., Wood-Kaczmar, A., Andre, R., Farshim, P., Bates, G. P., Holmans, P., Jones, L. and Tabrizi, S. J. ***Hum Mol Genet***, 2018 Oct 24. doi: 10.1093/hmg/ddy375.

* These authors should be regarded as joint first authors.

Chapter 7 Fan1 knockdown in R6/2 mice

7.1 Background

7.1.1 Genetic modifiers

Motor onset in Huntington's disease (HD) is inversely correlated with CAG repeat length, but still varies by several decades in patients with the same CAG repeat length, as measured in blood (Gusella et al., 2014, Keum et al., 2016). In human HD, the length of the *HTT* CAG repeat explains around 56% of variation in onset (Gusella et al., 2014, Langbehn et al., 2004), but about 40% of the remaining variability is heritable and due to genetic differences elsewhere in the genome (Wexler et al., 2004a). Investigating these genetic modifiers, the GeM genome-wide association study (GWAS) of HD age at onset (AAO) identified a locus likely underlain by the DNA interstrand cross link repair nuclease FAN1, a protein known to interact with MMR components such as MLH1 (GeM-HD, 2015). Two independent signals underlie this locus; the minor allele at rs146353869 was associated with 6.1 year earlier onset (p = 4.3E-20) and at rs2140734 with 1.4 year later onset (p = 7.1E-14). This suggests polymorphisms at this locus can be damaging or protective in HD. The third most significant SNP at the chromosome 15 locus encodes an amino acid change within *FAN1* (p.R507H, p = 9.34E-18) that is predicted *in silico* to be functionally deleterious. Pathway analysis showed DNA repair gene sets were associated with disease onset. Therefore, FAN1 may be part of a of DNA damage response (DDR) network that modulates HD pathogenesis (GeM-HD, 2015).

Bettencourt et al. (2016) found that some of the most significant DNA repair variants from this GeM GWAS, including those in *FAN1*, *RRM2B* and MMR component *PMS2*, also influence onset in other CAG repeat expansion diseases, suggesting DNA repair is a common mechanism driving severity of numerous diseases (Chapter 3). A recent GWAS found that genetic variation in *MSH3* was associated with slow HD progression (Chapter 8) (Hensman Moss et al., 2017b). The second most significant association region once again tags the chromosome 15 locus likely underlain by *FAN1* (p=2·35E-06).

7.1.2 Repeat instability

In HD, the pathogenic CAG repeat is unstable, tending to expand over time, particularly in the striatum, the tissue most prominently affected neuropathologically, correlates with onset (Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009) and likely contributes to disease progression. In HD transgenic mice, the CAG repeat expands most markedly in striatal neurons, correlating with phenotypic severity. It also expands in the cortex and liver, but is stable in the cerebellum, blood and tail (Gonitel et al., 2008, Lee et al., 2011a, Mangiarini et al., 1997). As this expansion occurs in postmitotic neurons, continues when the cell cycle is arrested (Gomes-Pereira et al., 2014b), and does not occur in mice when the HD transgene is not expressed (Mangiarini et al., 1997), expansion likely occurs independent of DNA replication.

When expansion occurs in germ cells, particularly prominent through the paternal line in HD, it results in intergenerational expansion and earlier onset in successive generations, a phenomenon known as anticipation. There is significant CAG length mosaicism in HD patients' sperm, which correlates with expansion on transmission (Telenius et al., 1995). In transgenic mice, expansion occurs after meiosis, again implicating DNA repair or transcription rather than replication (Kovtun and McMurray, 2001).

Recent evidence demonstrates that Fan1 protects against expansion of the CGG repeat tract in a mouse model of Fragile X (Zhao and Usdin, 2018). A similar stabilisation of the *HTT* CAG repeat tract would reduce somatic instability and could underlie FAN1's effect on HD course.

7.1.3 FAN1

FAN1 is a DNA endo- and exonuclease originally described by four groups in 2010 (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Smogorzewska et al., 2010a). It is required for interstrand crosslink repair (ICL), though its precise role in this process remains unclear (Thongthip et al., 2016, Lachaud et al., 2016a, Lachaud et al., 2016b). Repair occurs in complex with at least some mismatch repair (MMR) proteins (MacKay et al., 2010b, Kratz et al., 2010a, Liu et al., 2010b, Smogorzewska et al., 2010a). FAN1 is structure rather than sequence specific, binding branched DNA structures that mimic DNA repair intermediates (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b) and cleaving at the 5' flap (Pennell et al., 2014). It also has an independent role maintaining genomic stability and preventing chromosomal abnormalities, possibly through the regulation of replication fork dynamics (Lachaud et al., 2016a, Chaudhury et al., 2014).

It has four characterised domains. Through its ubiquitin binding domain (UBZ) it interacts with monoubiquitinated FANCD2 and FANCI of the FA pathway of ICL repair, which are involved in localisation to nuclear ICL damage foci (Liu et al., 2010c, Smogorzewska et al., 2010a). Its DNA binding (SAP) domain is structure rather than sequence specific, binding branched DNA structures that mimic DNA repair intermediates (Kratz et al., 2010a, Liu et al., 2010c, MacKay et al., 2010a), and may also be involved in recruiting FAN1 to ICL damage foci (Thongthip et al., 2016). The tetratricopeptide repeat (TPR) mediates protein-protein interactions and the assembly of multiprotein complexes. Finally, its nuclease domain, a viral replication and repair nuclease (VRR Nuc), has endonuclease activity at 5' flap structures and is a 5'-3' exonuclease (MacKay et al., 2010b). FAN1's crystal structure has been determined bound to DNA substrates and suggests it may form a dimer to orient and nick DNA (Wang et al., 2014b, Zhao et al., 2014, Gwon et al., 2014, Yan et al., 2015).

FAN1 cleaves DNA at interstrand crosslinks (ICL) and repair occurs in complex with at least some DNA repair proteins, including the ID complex (FANCD2 and FANCI) of ICL repair and mismatch repair complexes MutLα (MLH1 and PMS2) and MutLγ (MLH1 and MLH3) (MacKay et al., 2010a, Kratz et al., 2010a, Liu et al., 2010c, Smogorzewska et al., 2010b). Its interaction with components of MMR system provides a plausible functional link between our genetic data (GeM-HD, 2015, Bettencourt et al., 2016) and a role in somatic instability.

Unlike other FA genes, *FAN1* mutations do not cause Fanconi anaemia. However, loss of function mutations cause karyomegalic interstitial nephritis, a recessive renal syndrome (Zhou et al., 2012, Lachaud et al., 2016b, Thongthip et al., 2016), and heterozygous truncating mutations, like mutations in MMR proteins, cause pancreatic (Smith et al., 2016) and hereditary colorectal cancers (Segui et al., 2015b). *FAN1* variants may also be associated with susceptibility to schizophrenia and autism (Ionita-Laza et al., 2014).

7.2 Aims

Mismatch repair proteins may act on abnormal structures formed by CAG repeat DNA, such as hairpin loops, attempting repair and resulting in expansion. FAN1 may bind at or near the CAG repeat and, through its interaction with DNA damage response components, protect against CAG expansion. This chapter aims to assess the impact of reducing *Fan1* expression by AAV-delivered miRNA on somatic instability in the striatum and liver of R6/2 mice.

7.3 Methods

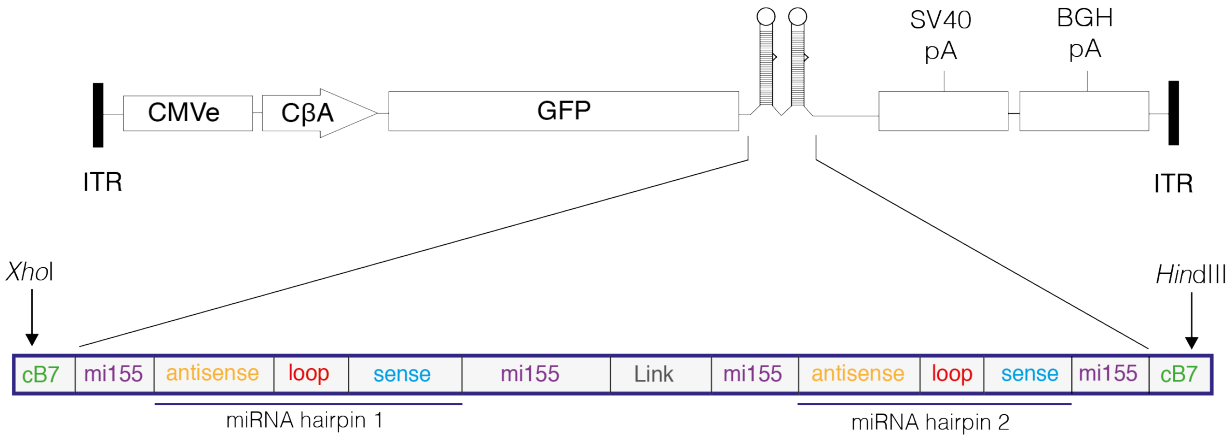
7.3.1 Viral vector

7.3.1.1 Virus

Adeno-associated virus (AAV) is a single-stranded DNA virus that can produce long term transduction of both dividing and non-dividing cells. They are preferred over lentiviruses as they remain primarily episomal, rather than integrating into the genome, so do not disrupt chromatin at the site of integration which can affect transgene and neighbouring gene expression. AAV infections produce only mild immune responses and are considered non-pathogenic (Aschauer et al., 2013). Serotypes 1-9 have been extensively used for gene therapy. Capsid protein is a major determinant of cellular tropism and transduction efficiency (Van Vliet et al., 2008), so hybrid vectors have been engineered using the genome of serotype 2 and capsid proteins of serotypes 1-9. AAV2/9 was selected for its efficient transduction of mouse striatum, hippocampus and cortex (Aschauer et al., 2013, Limberis and Wilson, 2006). Our group has previousl found widespread expression of a GFP reporter across the brain when delivered by stereotactic striatal injection.

7.3.1.2 Target sequence

Two *Fan1* target sequences were selected, based on commercially available siRNA oligonucleotides (Dharmacon) previously used in our lab to successfully reduce *Fan1* expression *in vitro* in mouse embryonic fibroblast (MEF) cells. Artificial miRNA design incorporating these target sequences was based on the Block-IT polII system (Invitrogen). Target sequences are 19mers, so a base was added at each end of the *Fan1* sequence to make 21mers. Bases 9 and 10 were removed from the sense sequence to produce a bulge in the hairpin structure that mimics that of endogenous miR155 and is important for its function as an artificial miRNA. The link between the hairpins has an A>T substitution to remove the HindIII site (underlined in the sequences below).



Oligo 09:

CTAGACTCGAGGACGGGGTGACTGGAGGCTTGCTGAAGGCTGTATGCTGTTAAGTCGGAGGCAATCTCTTGTTTTGGCCAC
TGACTGACAAGAGATTCTCCGACTTAACAGGACACAAGGCCTGTTACTAGCACTCACATGGAACAAATGGCCCtagcttcc
cgggataggtacCTGGAGGCTTGCTGAAGGCTGTATGCTGTTAAGTCGGAGGCAATCTCTTGTTTTGGCCACTTGACTGAC
AAGAGATTCTCCGACTTAACAGGACACAAGGCCTGTTACTAGCACTCACATGGAACAAATGGCCCACTACGCCTGAATCAA
GCTTATC

Oligo 10:

CTAGACTCGAGGACGGGGTGACTGGAGGCTTGCTGAAGGCTGTATGCTGGTAATCGAATGACACTGGCTTGTTTTGGCCAC
TGACTGACAAGCCAGTCATTCGATTACCAGGACACAAGGCCTGTTACTAGCACTCACATGGAACAAATGGCCCtagcttcc
cgggataggtacCTGGAGGCTTGCTGAAGGCTGTATGCTGGTAATCGAATGACACTGGCTTGTTTTGGCCACTTGACTGAC
AAGCCAGTCATTCGATTACCAGGACACAAGGCCTGTTACTAGCACTCACATGGAACAAATGGCCCACTACGCCTGAATCAA
GCTTATC

Figure 7.1. miRNA design.

Top *schematic representation of miRNA construct.* ***Bottom*** *– sequences of the two miRNA oligonucleotides used in the present study. Colours on the sequences and schematic are matched.*

7.3.1.3 Cloning

The sequences were synthesised by GeneArt and cloned into the AAV9 cB7 GFP vector, 3’ of the

open reading frame (ORF) as tandem amiRNAs by Penn Vector Core, University of Pennsylvania. These paradigms were developed by Dr Sena-Esteves (University of Massachusetts), who advised on this phase of the project. Insertion in the 3' untranslated region, between the GFP stop codon and the WPRE, allows the production of one transcript encoding both GFP and our amiRNA molecules. This strategy was adopted to avoid potential *in vivo* toxicity associated with constitutive U6 RNA polymerase III driven shRNA transcription, previously shown to overload miRNA processing factors. It is expressed under the control of the CAGG promoter. Two sequences (oligo 09 and 10) and a scrambled control were cloned into the AAV9 cB7 GFP vector backbone. Knockdown of *Fan1* was tested by transient transfection of mouse fibroblast (3T3) cells.

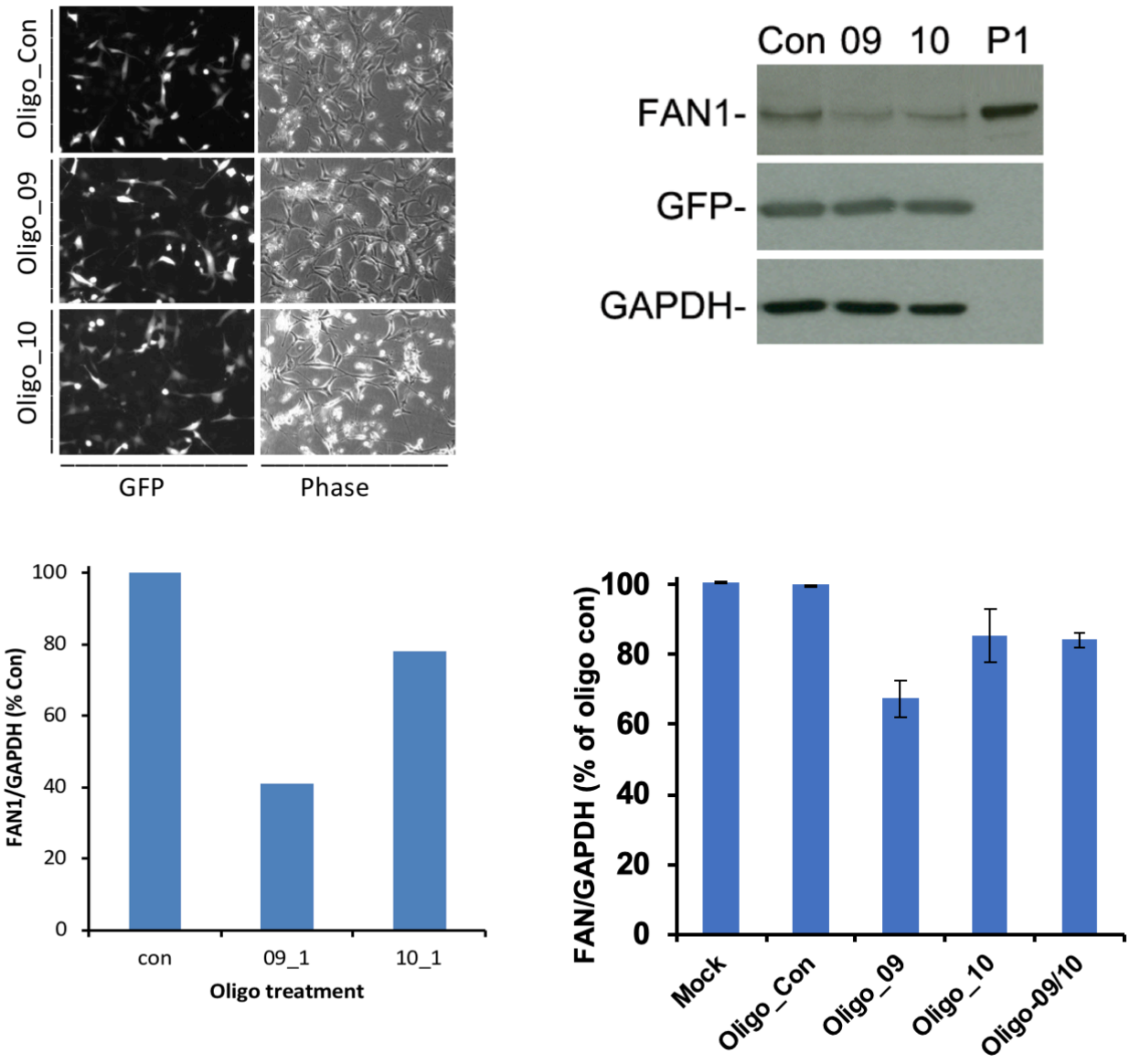


Figure 7.2. Fan1 silencing in 3T3 mouse embryonic fibroblasts (MEF) using AAV9 cB7 eGFP miRNA constructs.
Top left *– live cell GFP imaging of transfected cells with oligonucleotide 09, 10 or scrambled control.* ***Top right*** *– Immunoblot of samples probed with S101D anti-Fan1 (1:100 overnight, 5% DMP), anti-GFP and anti-GAPDH. P1 is a sample prepared from mouse brain.* ***Bottom left*** *– Densitometric quantification of the immunoblot demonstrating 60% and 20% Fan1 knockdown with oligonucleotides 09 and 10 respectively.* ***Bottom right*** *– mean Fan1 knockdown from five transfections. Error bars represent SEM. Note approximately 35% knockdown with oligonucleotide 09 relative to scrambled control.*

Oligonucleotide 09 and a scrambled control sequence were selected for viral packaging.

- AAV2/9.CB7.CI.eGFP-miR.mFan1.WPRE.rBG (3.53 x 1013 genome copies/ml ddPCR)
- AAV2/9.CB7.CI.eGFP-miR.Control.WPRE.rBG (3.78 x 1013 genome copies/ml ddPCR)

7.3.2 Experimental groups

7.3.2.1 Toxicity study

To investigate whether there are any adverse effects associated with *Fan1* knockdown (KD) at 4 weeks age, the AAV2/9 vector was administered by four regimens.

1. Direct unilateral intrastriatal (IS) injection into the left striatum
2. IS injection and intravenous (IV) tail vein administration
3. IV tail vein route only
4. Intraperitoneal (IP) route only

For these experiments, the R6/2 colony was maintained by backcrossing to (CBA/Ca x C57BL/6J)F1 mice. Animals were monitored for 4-5 weeks post-injection for signs of pain, weight loss (at least twice weekly) and routine health checks for subdued behaviour, hunched appearance, piloerection, seizures, difficulty handling or jumpy behaviour, nose bulge or swollen cheeks, wound scratching or weight loss of greater than 15%. A total of 24 animals were used, aiming to keep numbers to a minimum. Treated mice were compared to a control, uninjected group. The IP injection protocol was performed on *Hdh*Q150 knock-in colony mice on a C57BL/6J background.

Treatment group	Genotype	Age (wk)	Gender (M/F)	CAG repeat length (sd)	Treatment date	Dissection date
Intrastriatal injection only	WT	4	2/2	-	09/05/2017	13/06/2017
	R6/2	4	3/1	181 ± 2.13	09/05/2017	13/06/2017
Intrastriatal and tail vein injections	WT	4	2/2	-	10/05/2017	13/06/2017
	R6/2	4	1/3	182 ± 0.32	10/05/2017	13/06/2017
Tail vein only	WT	4	2/2	-	11/05/2017	13/06/2017
	R6/2	4	2/2	181 ± 1	11/05/2017	13/06/2017
Intraperitoneal injection only	WT	4	1/3	-	06/06/2017	04/07/2017
	Het HdhQ150	1	0/1	-	06/06/2017	04/07/2017
Control (uninjected)	WT	4	-	-	-	13/06/2017
	R6/2	4	-	180 ± 1.55	-	13/06/2017

Table 7.1. Animals used in the toxicity study.

7.3.2.2 Experimental study

Fan1 miRNA or scrambled miRNA (control) virus was delivered to R6/2 mice by combined left striatal and intraperitoneal (IP) injection, the optimal protocols selected following toxicity experiments. Mice from a (CBA/Ca x C57BL/6J)F1 colony were backcrossed to C57BL/6J for one generation. A total of 80 mice were used. Weights and temperature were measured at least twice weekly. They were co-housed with WT. Due to surgical time-constraints, mice in each group were injected over two days.

Study	Treatment	Treatment date	Age at treatment (wk)	Genotype	Gender [M/F] (# died in brackets)	Total	Age at dissection (wk)
8wk	Fan1 miRNA	20/06/2017	4	R6/2	3/3	6	8
		22/06/2017	4	R6/2	3/3	6	8
	Scr miRNA	20/06/2017	4	R6/2	3/2	5	8
		22/06/2017	4	R6/2	3/2	5	8
11wk	Fan1 miRNA	28/06/2017	4	R6/2	2(1)/3	6(4)	11
		29/06/2017	4	R6/2	3/3(1)	6(5)	11
	Scr miRNA	28/06/2017	4	R6/2	3(1)/3(1)	6(4)	11
		29/06/2017	4	R6/2	3(1)/3(1)	6(4)	11
	Untreated control	-	4	R6/2	3/8	11	11
		-	4	WT	9/5	14	11
Immunohistochemistry study	Fan1 (perfuse)	11/07/2017	4	R6/2	2/2	4	8
		11/07/2017	4	WT	2/2	4	8
	Control (perfuse)	12/07/2017	4	R6/2	2/2	4	8
		12/07/2017	4	WT	2/2	4	8

Table 7.2. Animals used in experimental study.

7.3.3 Pilot studies

Pilot studies were conducted to determine the baseline level of *Fan1* expression across different tissues in R6/2 mice and to optimise methods for extraction of DNA, RNA and protein.

7.3.4 Surgical procedures

7.3.4.1 Stereotactic microinjection

The AAV2/9.mFan1 vector was used for toxicity studies. Viral preparations were delivered by unilateral intrastriatal stereotactic injection at a volume of 3 µL and at a rate of 0.1 µL/min for 5 min. The needle was left in place for 5 minutes before removing the injector. The stereotactic coordinates used to target the striatum were anteroposterior (Y): +0.5; mediolateral (X): +2.5 and dorsoventral (Z): -4.0 expressed in mm relative to the Bregma, according to the Allen mouse brain atlas. Animals were monitored after surgery.

7.3.4.2 Intravenous (IV) tail vein injection

5 µL of 3.53 x 1013 GC/mL (AAV2/9.Fan1miRNA) diluted in 100 µL of sterile PBS was administered to each animal using a 1 mL syringe with 30 G needle.

7.3.4.3 Intraperitoneal (IP) injection

5 µL of 3.53 x 1013 GC/mL (AAV9.Fan1miRNA) was diluted in 100µL of sterile PBS in a 1mL syringe with a 30 G needle and was administered to each animal

7.3.5 Experimental protocol

At either 4 weeks (8 week cohort) or 7 weeks (11 week cohort) after injection, the left striatum (injected), right striatum, rest of brain, liver and tail were snap frozen in liquid nitrogen and stored at -80°C. The protocol is outlined below.

Toxicity samples were collected at 8-9 weeks age to assess Fan1 protein levels and optimise qPCR assays.

Experimental samples were collected at 8 or 11 weeks, and assessed for *Fan1* knockdown by qPCR and Western blot, and studied for CAG repeat expansion.

Immunohistochemistry samples (fixed brain and liver) were collected at 8 weeks to assess for GFP expression.

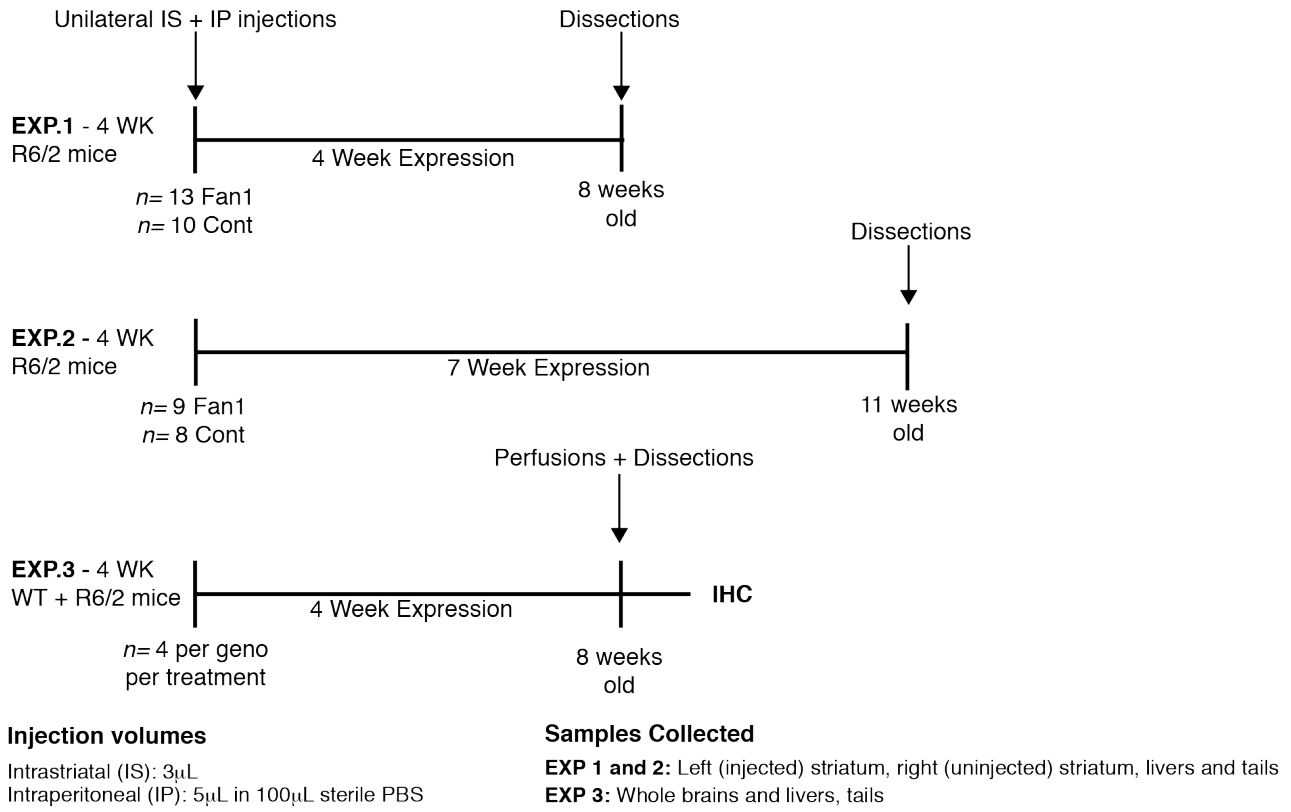


Figure 7.3. Experimental protocol.

7.3.6 Immunohistochemistry (IHC)

Mice were transcardially perfused 3 weeks after injection, first with heparinised PBS, and, then with 4% PFA solution. Whole brains and livers were harvested and post fixed in 4% PFA solution for an additional 24 hours. Fixed tissues were subjected to sucrose gradient (20% then 30% in PBS) and embedded in Tissue-Tek O.C.T and stored at -80°C. Coronal or sagittal brain sections (25 µm) were cut on a cryostat at 50 µm intervals and collected into 12-well plates containing tissue cryopreservative solution and stored in -20°C until immunostaining with an antibody to GFP (Invitrogen A11122 at 1:1000).

Coronal brain sections were assessed for GFP expression using a Nikon Eclipse A1R point scanning confocal microscope using 10x objective lens set to 8 x 8 fields stitching procedure with 15% overlap. Images were acquired in NIS-Elements AR Software. Representative images were then processed using Fiji (ImageJ Image Analysis) and prepared in Adobe Photoshop and InDesign CS6.

7.3.7 DNA extraction

DNA was extracted using a modified high salt method (Aljanabi and Martinez, 1997). Briefly, samples were lysed in 475 µL lysis buffer and 1 mg/ml proteinase K overnight at 50°C. 300 µL of saturated NaCl solution was added, then shaken vigorously and incubated for 2 min before centrifuging at full speed for 35 min. The supernatant, containing the DNA, was precipitated by adding to 650 µL of 100% ethanol, shaken and centrifuged at full speed for 20 min. The supernatant was discarded and the pellet resuspended in 300 µL of 70% ethanol before centrifugation at full speed for 5-15 min. The

supernatant was again discarded and the pellet dried at room temperature for an hour. The pellet was then resuspended in 50-150 μL of 5mM TRIS (adjusting buffer volume depending on pellet size, using 10 μL if no pellet was visible).

7.3.8 RNA extraction

Samples were stored at -80°C prior to lysis. Brain samples were lysed by homogeniser probe for 30 secs in 500 μL of Qiazol (Qiagen). Muscle and peripheral tissues were placed in ribolyser tubes, to which 700 μL of Qiazol was added. They were lysed using the Fast-Prep 24 (MP Biomedicals) at 6.5 m/s for 1 min three times. Chloroform (VWR) was added to lysed samples (200 μL for brain and 250 μL for peripheral tissues), which were then vortexed for 30 secs and centrifuged at 13,000 rpm for 15 min at room temperature. The aqueous phase was transferred to a new tube and an equal volume of 70% ethanol added. RNA was purified as per the RNeasy Mini Kit protocol (QIAGEN, 74106). A 15 min genomic DNA digestion step (DNAse I, QIAGEN, 79254) was performed between the RW1 buffer washes. RNA was eluted with water and concentration was measured on a NanoDrop 1000.

7.3.9 Protein extraction

Tissue was homogenised in RIPA buffer with protease inhibitors and benzonase using a small Eppendorf pestle until smooth. Protein concentration was determined by Bio-Rad assay. The tissue suspension was methanol precipitated, then resuspended in 1x SDS sample buffer to 4 mg/ml.

7.3.10 3-in-1 DNA, RNA and protein extraction

Extraction was performed according to the manufacturer's protocol (GE Healthcare, cat #28-9425-44). Briefly, samples were lysed in lysis buffer then added to a DNA column and centrifuged, saving the flow through which contains RNA and protein. The bound DNA was washed twice in wash buffer, then eluted in elution buffer. Acetone was added to the flow through, which was then added to the RNA column and centrifugation. RNA was treated with DNase I, washed with wash buffer and eluted in elution buffer. To the protein-containing flow through a protein precipitation buffer was added. The sample was vortexed, incubated at room temperature for 5 min, then centrifuged to remove supernatant. The protein pellet was suspended in water, before pelleting and removal of supernatant. The protein was then resuspended in a urea and detergent-containing 2-D DIGE buffer.

7.3.11 Quantitative real time PCR (qPCR)

The following Taqman probe sets were used.

Gene	Gene name	TaqMan probe (ThermoFisher)
Atp5b	ATP synthase subunit 5b	Mm01160389_g1
Eif4a2	Eukaryotic translation initiation factor 4A2	Mm01730183_gH
Fan1	FANCD2 And FANCI Associated Nuclease 1	Mm00625959_m1
Gapdh	Glyceraldehyde-3-phosphate dehydrogenase	Mm99999915_g1
Rpl13a	Ribosomal protein L13a	Mm05910660_g1
Sdha	Succinate dehydrogenase complex flavoprotein subunit a	Mm01352366_m1
Ubc	Ubiquitin C	Mm02525934_g1

Table 7.3. TaqMan qPCR probes.

7.3.12 Western blot

60 μg of each protein sample was loaded on a 9% gel, then transferred for 2 hours at 100 V. For immunoblotting, the gel blot was blocked in 10% dimethyl pimelimidate (DMP) for 1 hour, before the addition of the indicated primary antibody

(see below) overnight at 4°C. Secondary antibodies were added in Li-Cor buffer and incubated at room temperature for 90 min. The blot was scanned using a Li-Cor Odyssey.

Fan1:

- Primary antibody – Mouse Fan1 S101D, 1:250
- Secondary antibody – Goat anti-mouse, 1:4000

7.4 Contributions

This study was conceived and designed by Professors Bates and Tabrizi. Mice were bred and maintained in Professor Bates' lab. Viral injection and immunohistochemistry were performed by Pamela Farshim (UCL). The viral vector was designed by Rob Goold (UCL), with the advice of Dr Sena-Esteves (University of Massachusetts), and cloned by Penn Vector Core. Optimisation of extraction assays was performed by Michael Flower, with assistance from Rachel Flomen, Nadira Ali and Rob Goold (UCL). Tissue preparation was performed by Michael Flower. Quantative PCR was performed by Michael Flower and Nadira Ali. CAG repeat sizing was performed by Michael Flower and Rachel Flomen. Western blots were performed by Rob Goold. Data analysis was conducted by Michael Flower.

7.5 Results

7.5.1 Pilot studies

7.5.1.1 Fan1 expression in R6/2 tissues

In a pilot study, *Fan1* expression was measured in different tissues from two 4 week old and two 14 week old R6/2 mice by real time qPCR. Housekeeping gene expression was consistent across tissues. Note only one animal was available for 4 wk cord and no data is available for 14 wk hippocampus or 4 wk striatum. Amplification failed for *Sdha* in one animal for 14 wk adrenal and kidney tissue, so these animals were excluded from further analysis. *Fan1* is expressed at a relatively low level compared with housekeeping genes, as suggested by the relatively high cycle threshold values across all tissues.

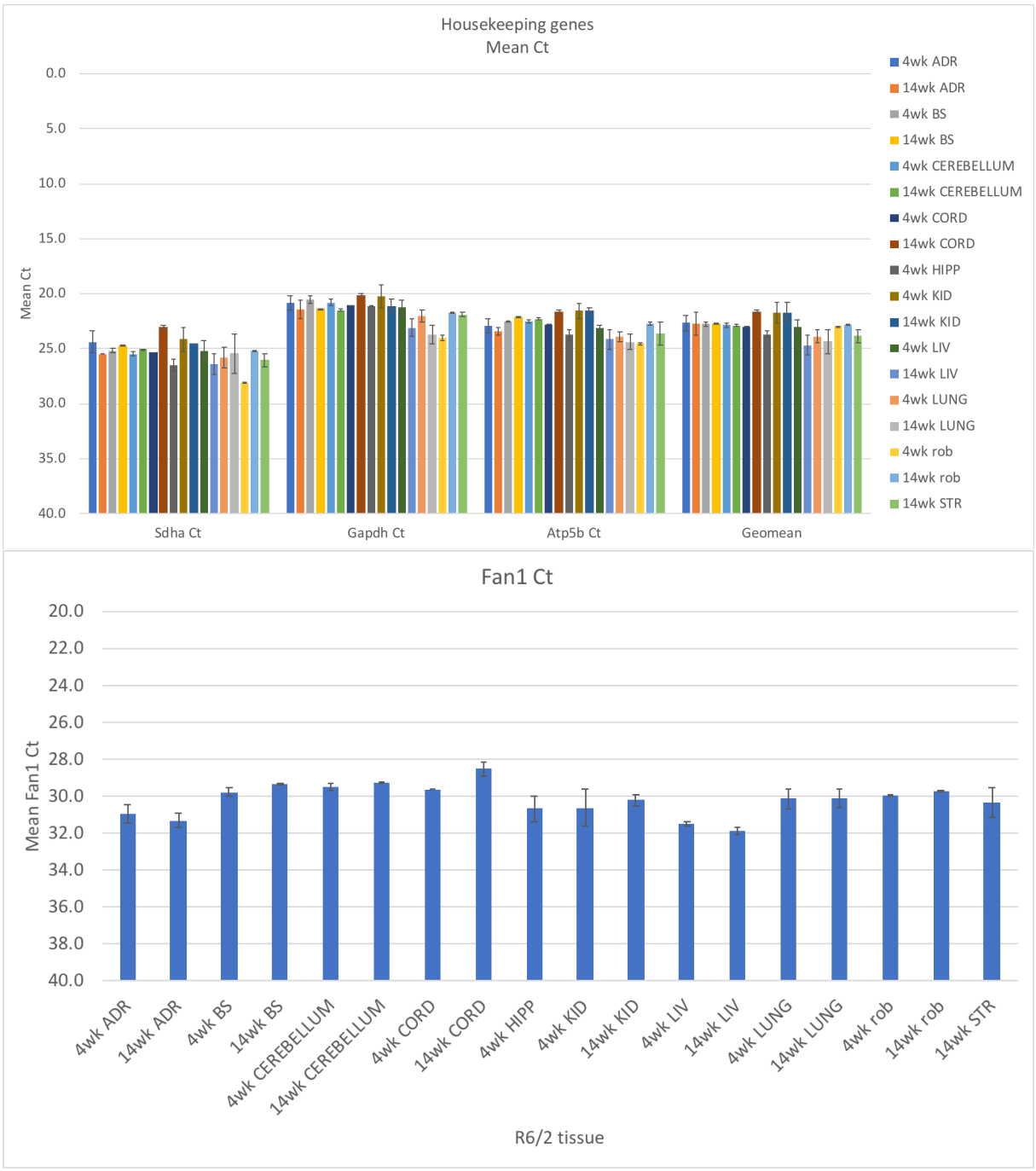


Figure 7.4. qPCR cycle threshold in pilot study of Fan1 expression in R6/2 tissues at 4 and 14 weeks.
Ct – mean cycle threshold (±sem). Note no tissue was analysed for 4 wk striatum and failure of amplification in one animal from the 14 wk adrenal and 14 wk kidney groups. ADR – adrenal, BS – brainstem, HIPP – hippocampus, KID – kidney, LIV – liver, rob – rest of brain, STR – striatum.

Fan1 was expressed at relatively high level in lung and in 4 week cerebellum, spinal cord and hippocampus, and at relatively low level in adrenal, kidney and liver tissue. *Fan1* expression appeared to decrease with age in cerebellum and spinal cord. However, this pilot study was limited by the small number of animals included.

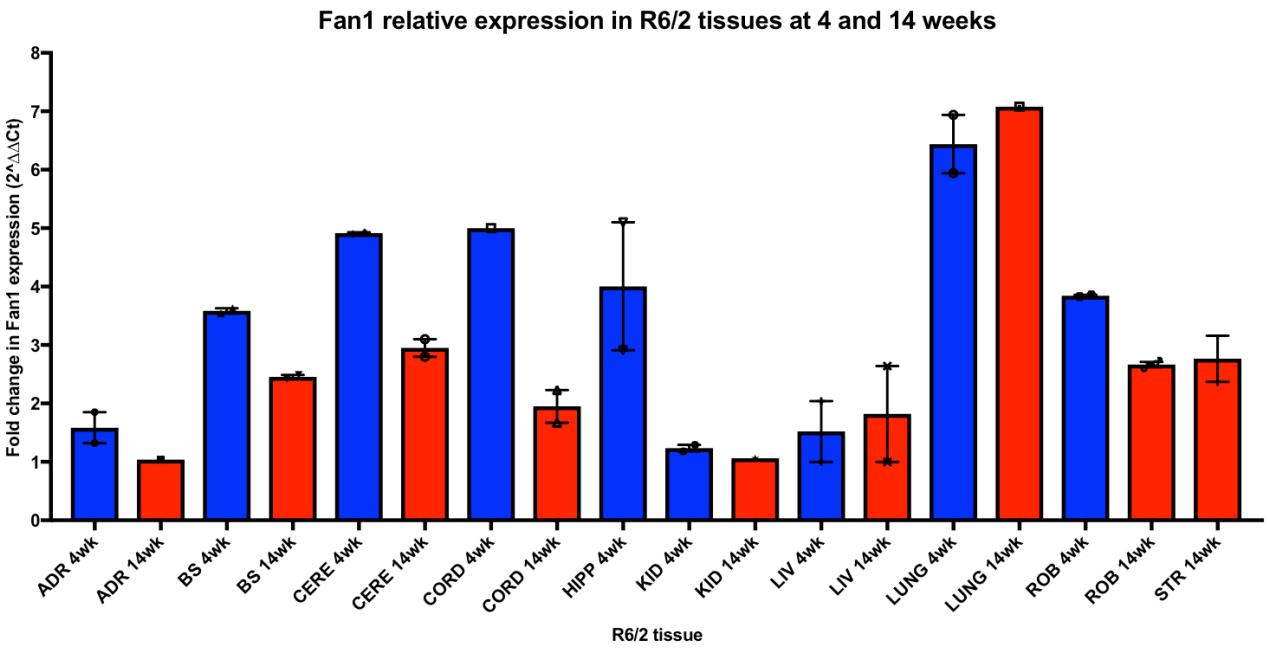


Figure 7.5. Relative Fan1 expression level in pilot study of R6/2 tissues at 4 and 14 wk age.
Blue – 4 week, red – 14 week. Note no tissue was analysed for 4 wk striatum and failure of amplification in one animal from the 14 wk adrenal and 14 wk kidney groups. ADR – adrenal, BS – brainstem, HIPP – hippocampus, KID – kidney, LIV – liver, ROB – rest of brain, STR – striatum.

7.5.1.2 Optimisation of assays

7.5.1.2.1 3-in-1 extraction of DNA, RNA and protein

The striatum is a small tissue from which samples were needed for fragment analysis (CAG repeat sizing), qPCR (*Fan1* transcript levels) and western blot (Fan1 protein expression). Before analysing the experimental samples, we determined which of the following methods provided optimal sensitivity and reliability.

1. 3-in-1 kit to extract DNA, RNA and protein from the same sample (GE Healthcare, #28-9425-44).
2. One third of each striatum delivered to each analysis method separately.

7.5.1.2.1.1 Protocol

Four 9 mg samples of cortex were taken from four R6/2 mice. This volume was selected to be equivalent to the mean weight of a left striatum. Each sample was delivered to either the 3-in-1 kit or traditional DNA, RNA or protein extraction, as detailed in Methods.

7.5.1.2.1.2 Results

DNA was analysed by nanodrop. The 3-in-1 and traditional DNA extractions gave equivalent purity, as demonstrated by the 260/280 and 260/230 nm ratios; nucleic acids have maximal absorbance at 260 nm, and absorbance at either 230 or 280 nm may suggest the presence of contaminants (Thermo). However, the DNA yield from the 3-in-1 kit was only 4% that of the traditional extraction method. For RNA extraction, the 3-in-1 kit gave a slightly lower purity than the traditional method and the total RNA yield was significantly reduced to 28%.

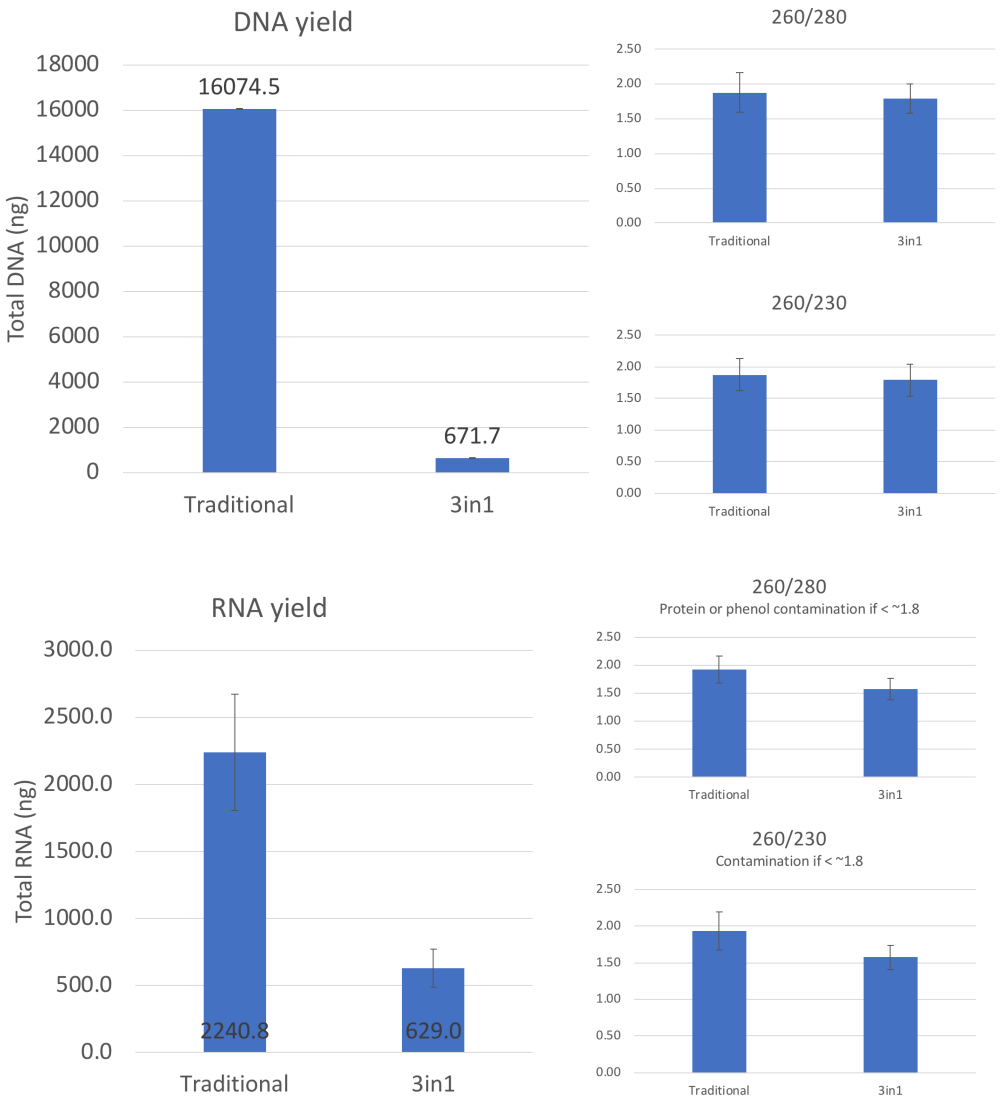


Figure 7.6. Comparing DNA and RNA yield and purity from 3-in-1 and traditional extractions.
***Top**– DNA, **bottom** – RNA. The left panel in each gives total DNA or RNA yield (concentration * elution volume). The right panels give absorbance ratios for 260/280 nm and 260/230 nm as indicated.*

qPCR showed the average cycle threshold was significantly higher for housekeeping genes *Atp5b* and *Sdha*, suggesting lower transcript levels in these eluted samples. As this affects only some housekeeping genes, certain transcripts appear to be lost disproportionately in the 3-in-1 extraction.

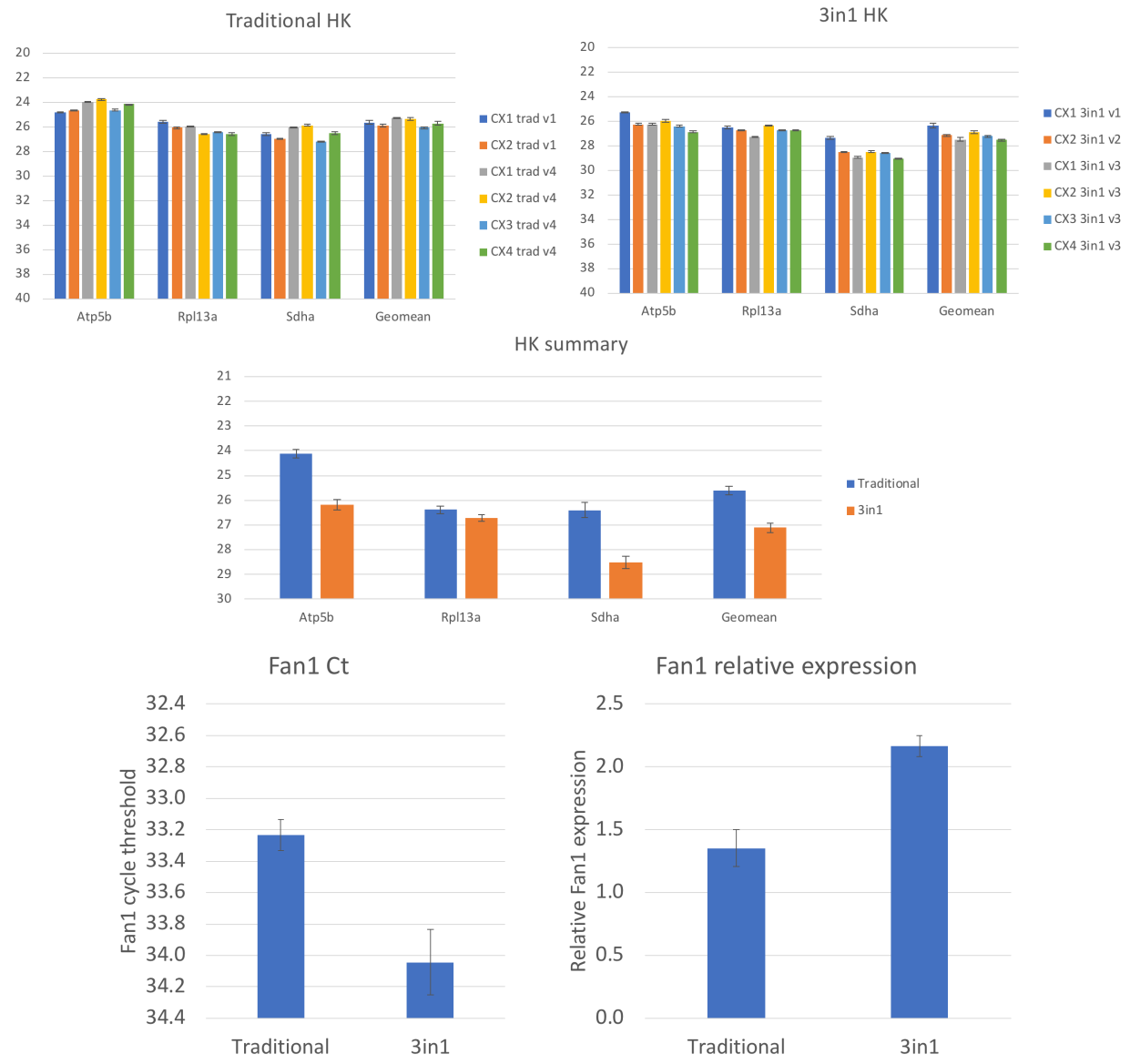


Figure 7.7. Comparing Fan1 expression level in 3-in-1 or traditional RNA-extracted cortex samples.
Ct – mean cycle threshold (±sem). ***Top left*** *– Ct for traditional RNA extraction.* ***Top right*** *– Ct for 3-in-1 RNA extraction.* ***Middle*** *– mean Ct for each housekeeping gene and geomean.* ***Bottom left*** *– mean Fan1 cycle threshold (±sem).* ***Bottom right*** *– mean relative expression, calculated by comparative Ct method (±sem).*

The cycle threshold for *Fan1* transcripts was also lower in 3-in-1 than traditional RNA extractions. However, relative expression, which controls for housekeeping gene expression level, suggested *Fan1* expression levels were higher in the 3-in-1 extracted samples. Once again, this suggests selective loss of some transcripts during the 3-in-1 extraction process.

Ponceau staining and western blot demonstrates significantly lower protein yield from the 3-in-1 extraction. Concerningly, there appeared to be specific loss of nuclear proteins such as Fan1 and histone H3, relative to Gapdh. This may because nuclear fractions are removed with DNA and RNA during 3-in-1.

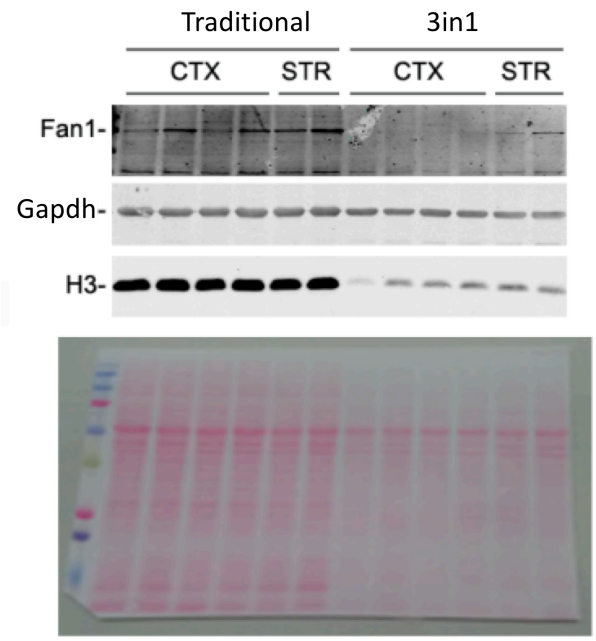


Figure 7.8. Western blot comparing Fan1 protein levels from 3-in-1 or traditional protein extractions from cortex.
Above *– western blot probed with antibodies to Fan1, Gapdh or histone H3.* ***Below*** *– Ponceau stained filter demonstrating protein bands.*

7.5.1.2.1.3 Conclusions

DNA, RNA and protein yields are significantly lower with the 3-in-1 kit compared to traditional extraction methods. Whilst DNA yield is adequate for fragment analysis, it appears there is selective loss of transcripts and nuclear proteins, which could lead to spurious results.

7.5.1.3 Traditional DNA, RNA and protein extraction from 1/3 striatum

Given the poor performance of the 3-in-1 extraction method, we next determined whether DNA, RNA and protein could reliably be extracted from 1/3 of a striatum.

7.5.1.3.1 Protocol

One striatum from each of 6 mice was divided into thirds, from which DNA, RNA or protein was extracted by traditional methods.

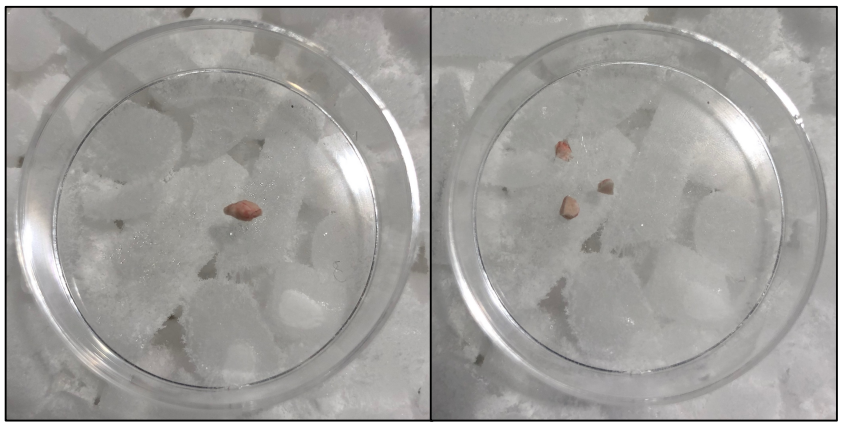


Figure 7.9. A striatal sample (left) divided into thirds (right).
Tissue was frozen on dry ice throughout the procedure.

7.5.1.3.2 Results

Fragment analysis was successfully performed on all samples. A representative trace is shown below, displayed in GeneMapper.

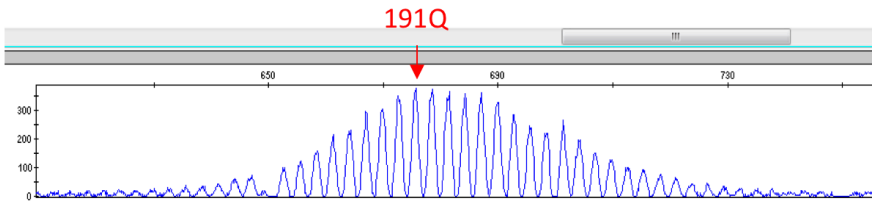


Figure 7.10. Fragment analysis from 1/3 of a striatum. Representative trace.

Protein was reliably extracted from 1/3 striatum, with a mean yield of 191 µg (95% CI 155-227 µg).

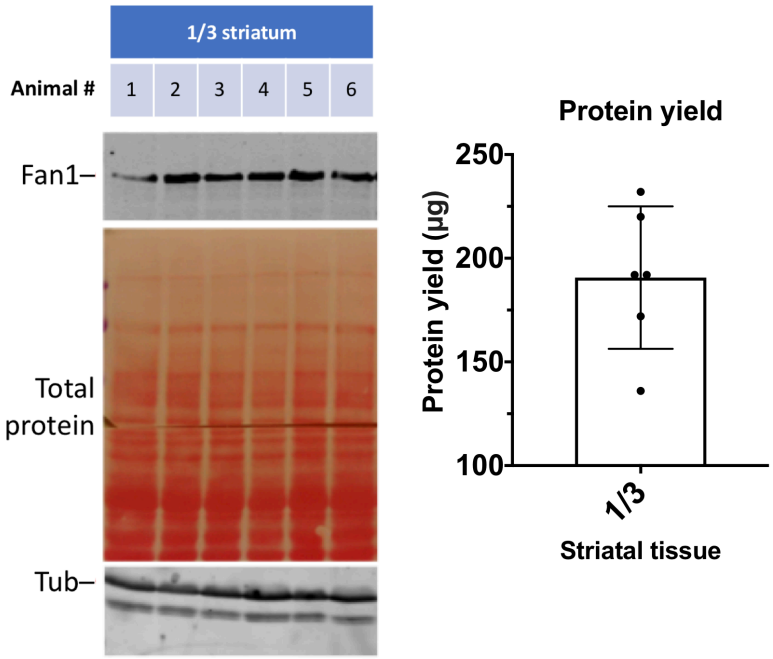


Figure 7.11. Protein extraction from 1/3 striatum of R6/2 mice.
***Left** – western blot with the antibodies indicated, in the middle is the Ponceau stained filter showing protein bands. **Right** – total protein yield (concentration * volume).*

RNA was reliably extracted from 1/3 striatum, giving a mean yield of 1340 ng (95% CI 894-1784 ng), consistent housekeeping gene levels and quantifiable *Fan1* expression.

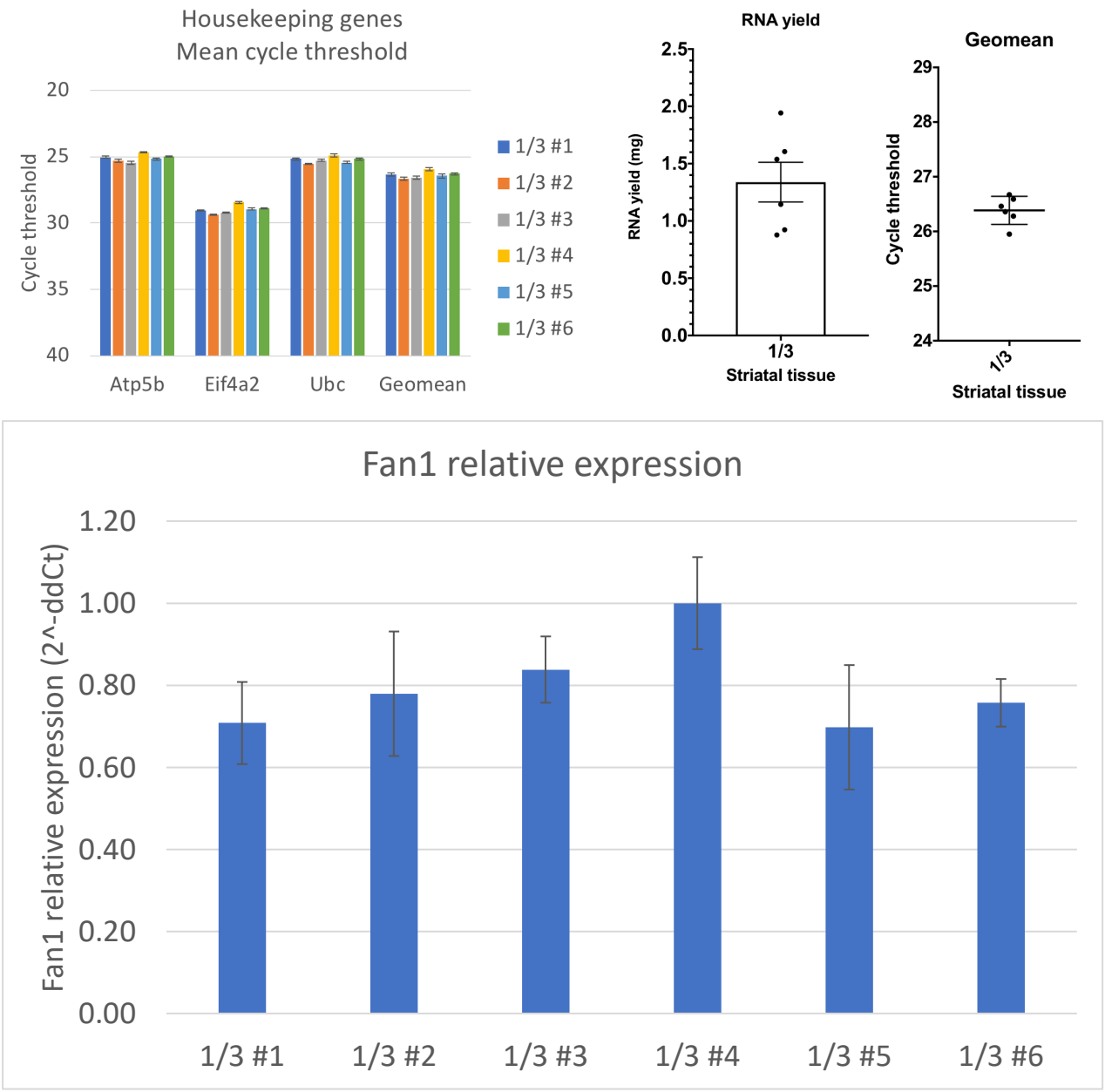


Figure 7.12. RNA extraction from 1/3 striatum and Fan1 expression.
Top left *– housekeeping gene cycle threshold.* ***Top middle*** *– RNA yield from 1/3 striatum.* ***Top right*** *– Geomean of housekeeping genes.* ***Bottom*** *– Fan1 relative expression in 1/3 striatum from six R6/2 mice.*

7.5.1.3.3 Conclusions

DNA, RNA and protein can all be reliably extracted from 1/3 of a striatum, permitting the analysis of CAG repeat length and Fan1 expression.

7.5.2 Toxicity study

There was no evidence of toxicity with any administration regimen.

7.5.2.1 Fan1 knockdown

WT and R6/2 animals treated by tail or striatal injection were compared. Lysates from striatum or liver were prepared and 40 µg probed with antibodies to Fan1, β3-tubulin or GFP. GFP expression was generally high in both genotypes, particularly on the left, which was directly injected. Fan1 levels appeared reduced compared to tail vein injected animals.

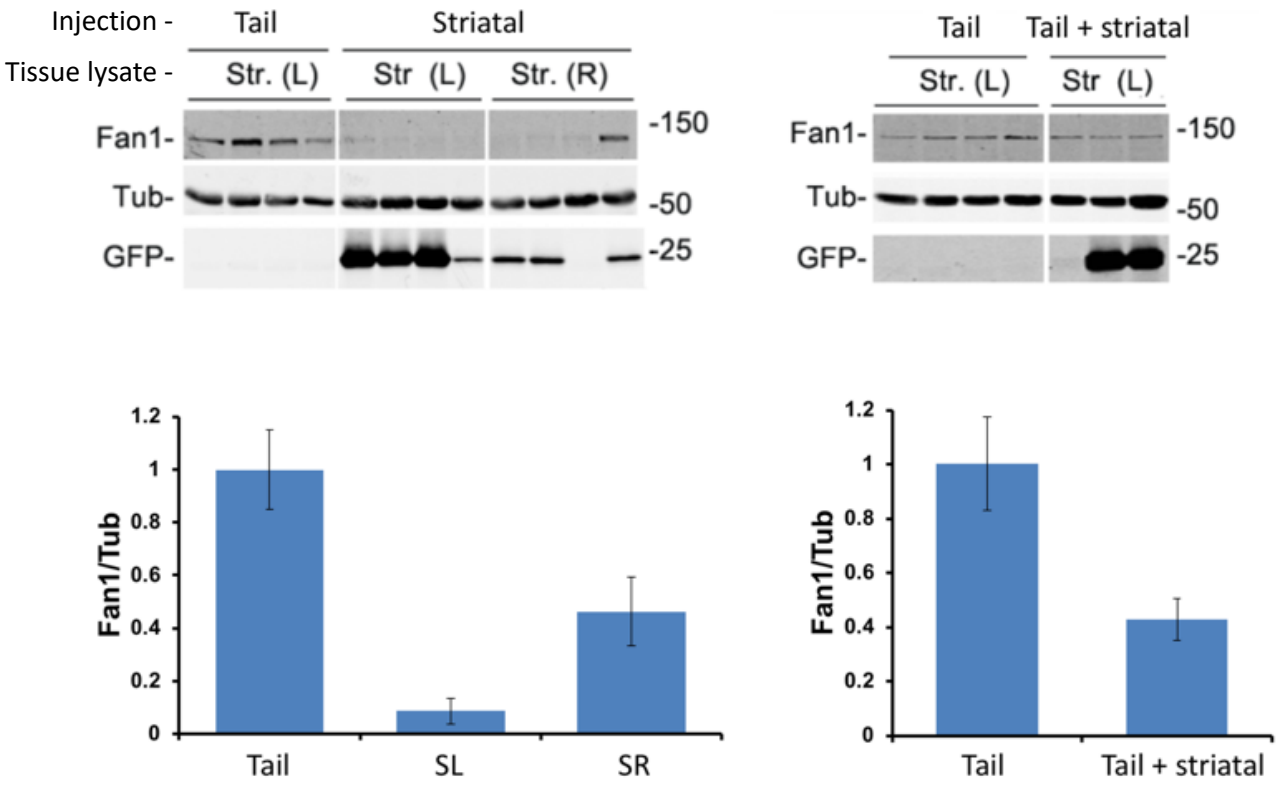


Figure 7.13. Fan1 knockdown in R6/2 following AAV9 cB7 eGFP.oligo 09 transduction.

Lysates were prepared from striatum of mice treated by tail vein (IV) or striatal injection. 40 µg was probed with the antibodies indicated. Left striatal injection produced significant Fan1 knockdown to around 8% of levels in intravenous treated animals (p = 4.96E-5).

In liver, GFP expression was lower than striatum, suggesting poorer transduction from both tail and striatal injections. *Fan1* expression in liver was very low or absent in liver in both treatment groups, independent of GFP expression level. This is consistent with published transcriptomic data which show *Fan1* is expressed at low levels endogenously (Papatheodorou et al., 2018).

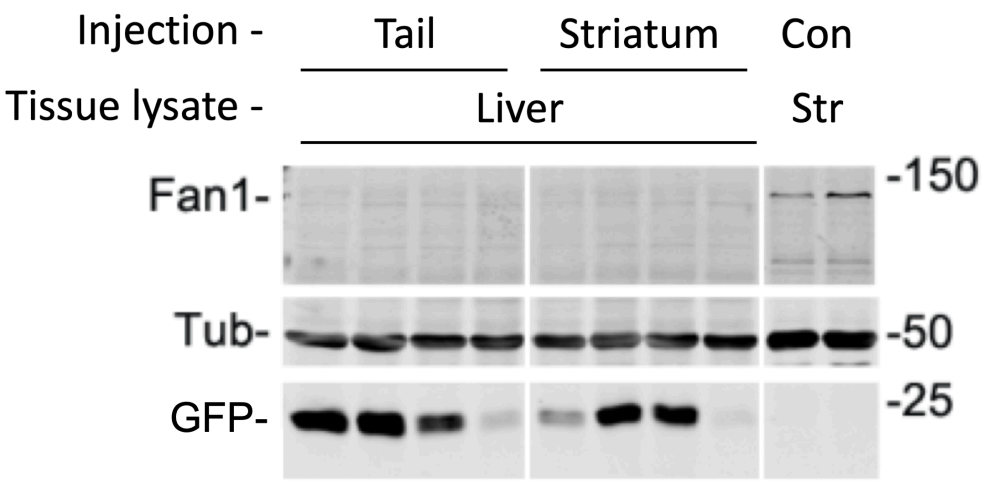


Figure 7.14. Fan1 knockdown in R6/2 liver.
Lysates were prepared from liver of mice treated by tail vein (IV, left) or striatal (middle) injection. 40 µg was probed with the antibodies indicated. Untreated striatal samples are loaded on the right for comparison. Note endogenous Fan1 levels are lower in liver than striatum.

7.5.3 Immunohistochemistry study

7.5.3.1 Striatum

Representative images showing CNS transduction following intrastriatal administration of either AAV2/9.CB7.CI.eGFP-miR.mFan1 (AAV9.mFan1) or AAV2/9.CB7.CI.eGFP-miR.Control vector (AAV9.Scramble).

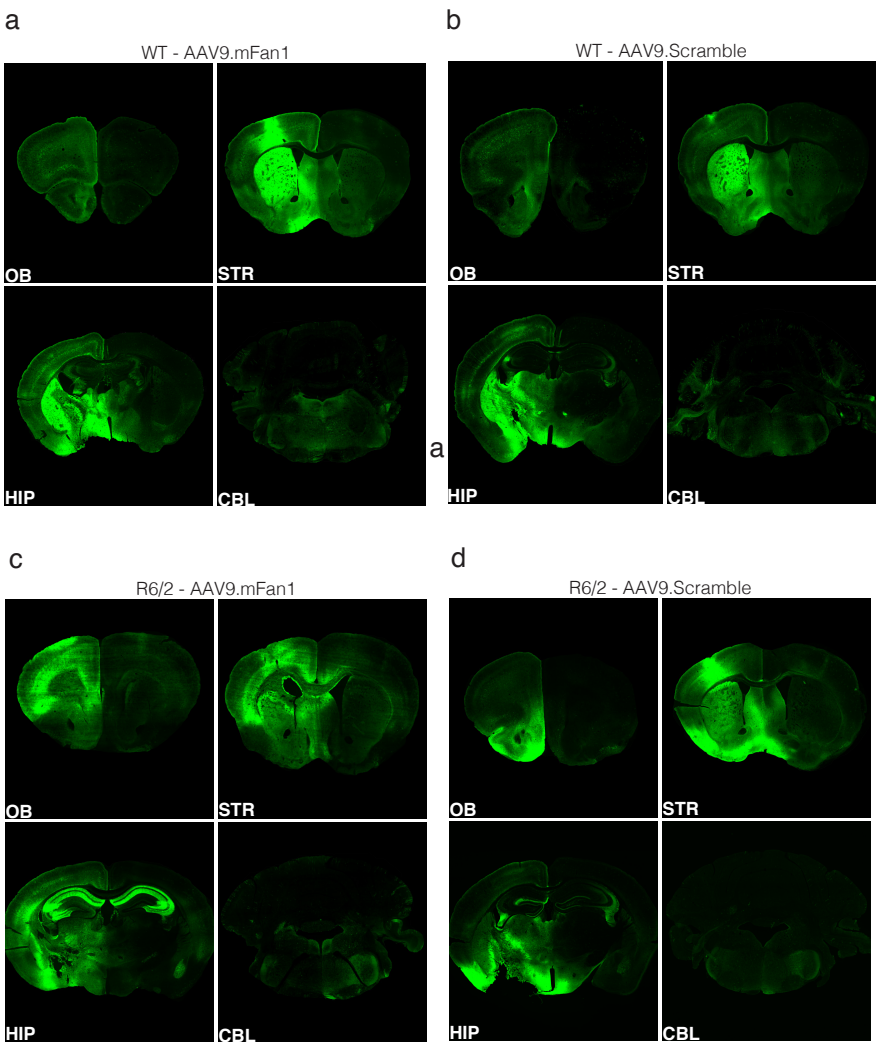


Figure 7.15. GFP expression following intrastriatal delivery of AAV9.mFan1 or AAV9.Scrambled control miRNA.
Confocal images of coronal sections taken at the level of the OB (olfactory bulb), STR (striatum), HIP (hippocampus) and CBL (cerebellum) in WT and R6/2 mice injected with AAV9.mFan1 (a and c) and WT and R6/2 mice injected with AAV9.Scramble miRNA vectors (b and d).

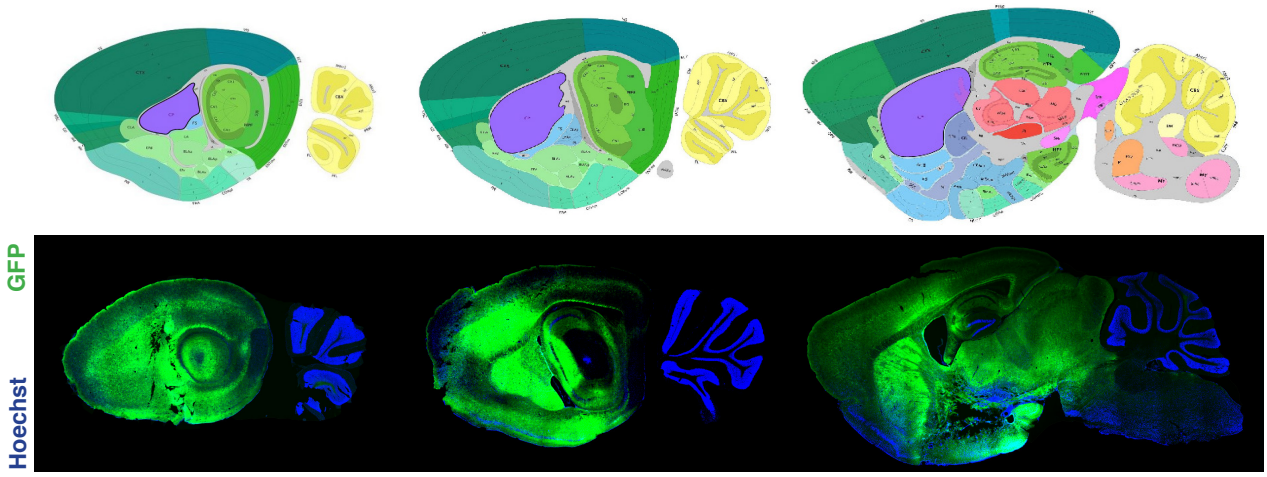


Figure 7.16. Representative sagittal sections showing GFP expression in the striatum.
Top diagrams taken from the Allen brain atlas, with caudate and putamen (CP) shown in purple.

The right side was also transduced by anterograde transport in R6/2 mice injected with AAV9.mFan1.miRNA (see below), though this was not observed in WT animals or with AAV9.scrambled.miRNA.

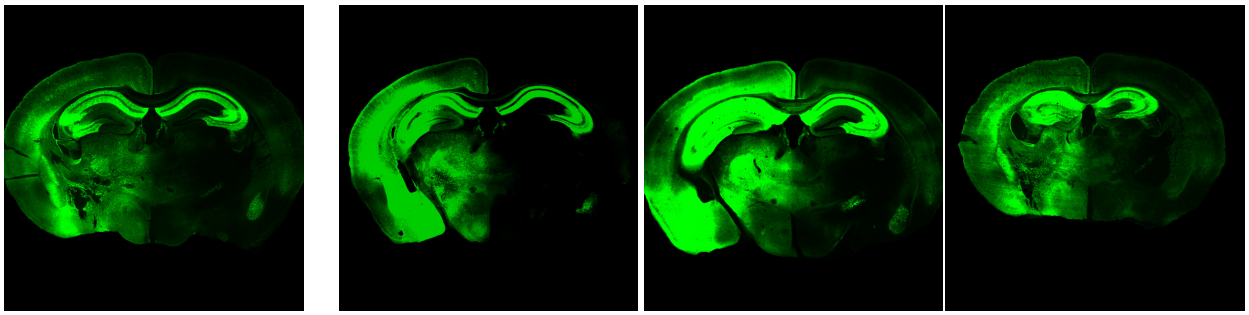


Figure 7.17. GFP distribution pattern in R6/2 mice receiving intrastriatal injection of AAV9.mFan1.miRNA.

7.5.3.2 Liver

GFP expression was assessed in liver sections from R6/2 and WT mice injected IP with AAV9.mFan1 and AAV9.Scramble miRNA vectors. The intensity of GFP signal was found to be lower in AAV9.Scrambled miRNA treated mice compared to AAV9.mFan1 treated WT and R6/2 mice.

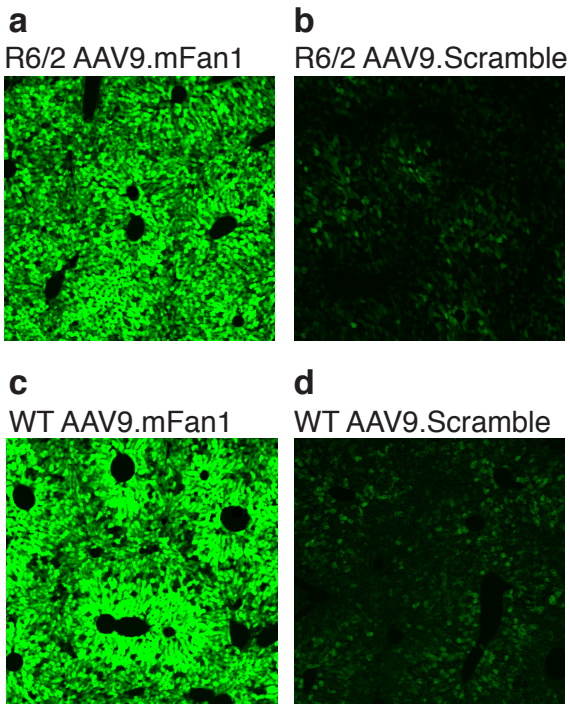


Figure 7.18. GFP expression in the liver.
*Representative sections from R6/2 mice treated with **a)** AAV9.mFan1, **b)** AAV9 scrambled miRNA vector and in WT mice treated with **c)** AAV9.mFan1 and **d)** AAV9.scramble miRNA vector.*

GFP intensity was lower is all animals receiving the AAV9.scrambled miRNA virus, compared with AAV9.mFan1, regardless of genotype.

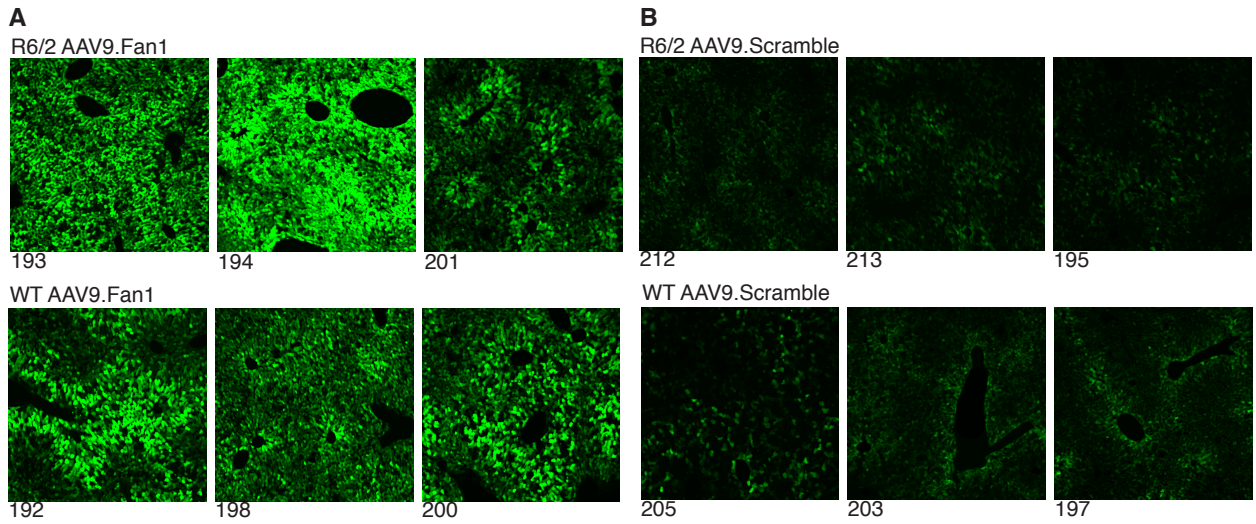


Figure 7.19. GFP expression pattern across all study groups receiving IP injections
*Injection of either **A)** AAV9.mFan1.miRNA or **B)** the AAV9.Scrambled control virus. Both R6/2 and WT mice that received the scrambled miRNA showed lower GFP intensity profiles as demonstrated in confocal images at x10 magnification. Numbers below each micrograph represent animal ID.*

GFP intensity was compared with the striatum. The lower intensity seen in scrambled liver was not seen in scrambled striatal sections.

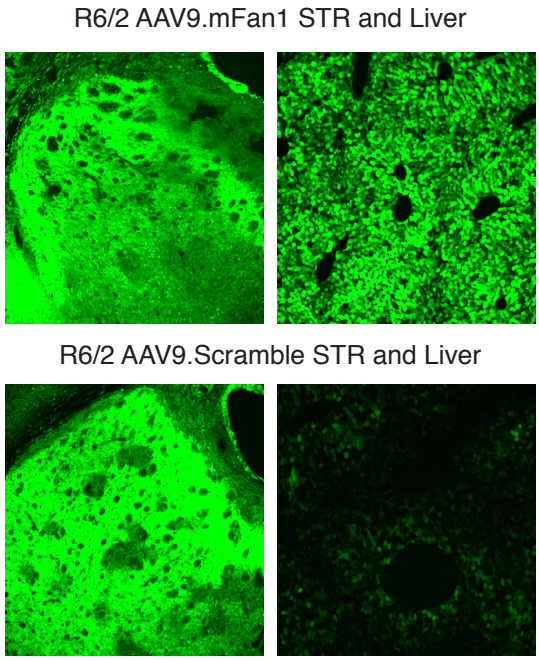


Figure 7.20. GFP intensity profile in transduced striatum and liver of R6/2 mice.
Striatum (left) and liver (right) of R6/2 mice receiving the AAV9.mFan1 miRNA (top) or AAV9.scrambled miRNA vector (bottom). No differences in GFP intensity were found in the striatum between the two treatment groups. Confocal images x10 magnification.

7.5.3.3 Mutant huntingtin aggregates

Analysis of striatal S830 antibody stained sections showed no difference in the number of mutant huntingtin aggregates between the Fan1 and scrambled miRNA-treated animals.

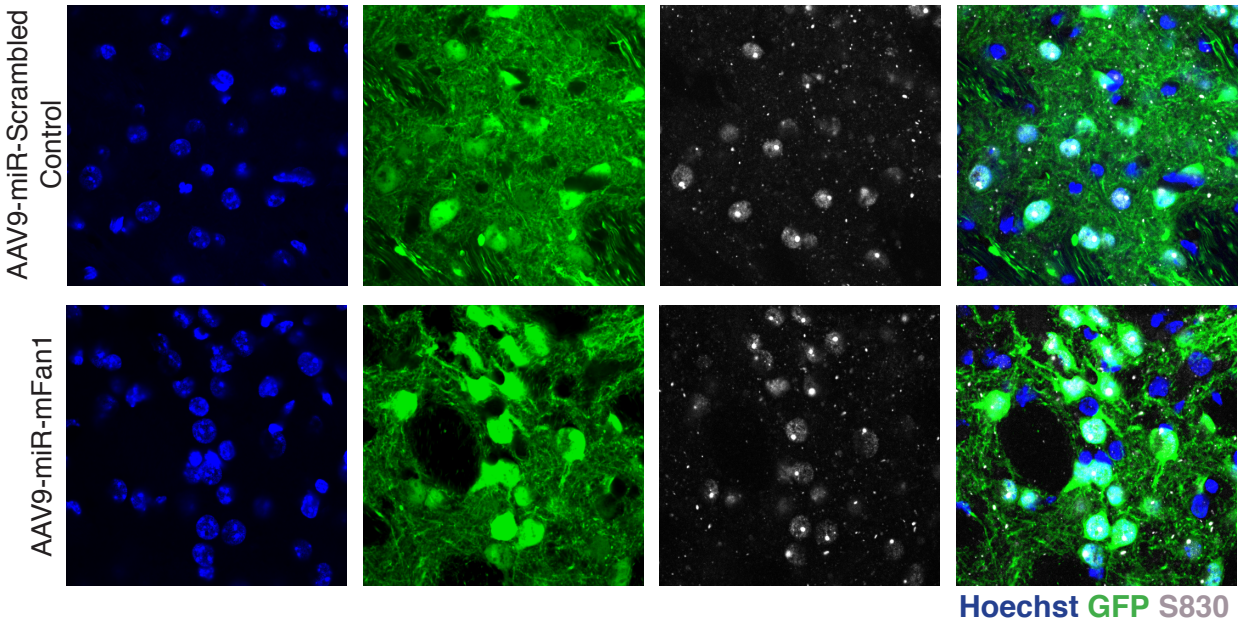


Figure 7.21. Mutant huntingtin aggregates in transduced R6/2 striatum.
Representative confocal images from the striatum of R6/2 mice transduced with AAV9.scrambled miRNA vector (top) or AAV9.mFan1 miRNA (bottom). S830 (white), GFP (virus), blue (Hoechst) at x100 magnification.

7.5.3.4 Conclusions

The left striatum of both R6/2 and WT mice was successfully transduced with both Fan1 and scrambled viruses. Successful transduction in the liver was achieved following IP delivery of AAV9.mFan1.miRNA, but GFP expression in the liver was lower for the scrambled control, regardless of genotype. AAV-mediated *Fan1* miRNA did not affect mutant huntingtin aggregation as demonstrated by S830 staining in striatal sections.

Though the combination of intrastriatal and IP was not formally assessed, these data suggested that this combination would have no adverse effects. Therefore, the **combination of intrastriatal and IP delivery** was used for the experimental study to optimise knockdown, decrease total procedure time, and reduce stress associated with restraint for tail vein injection.

7.5.4 Experimental study

7.5.4.1 Weight and body temperature

Animals were measured at least twice weekly throughout the duration of vector expression.

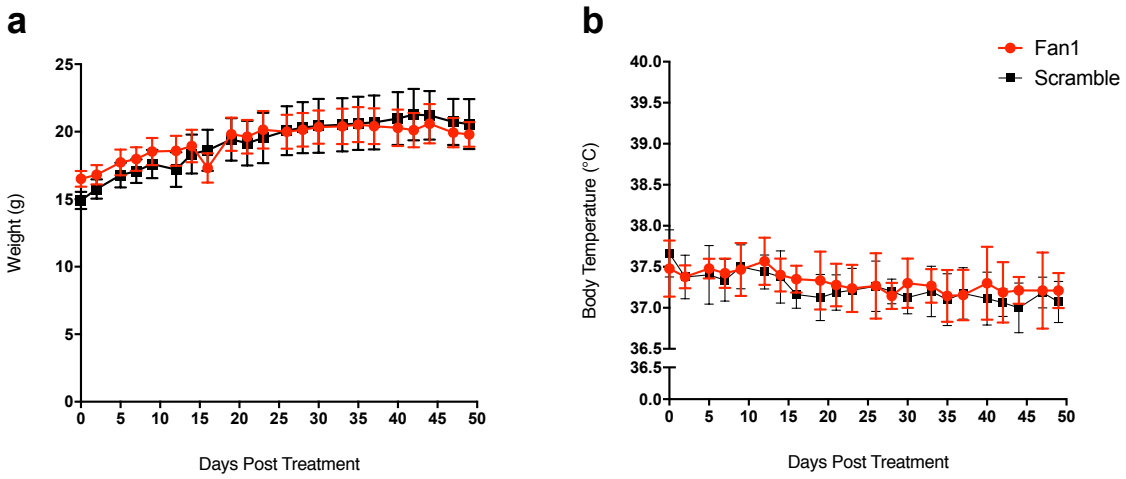


Figure 7.22. Weight and temperature in 11-week mice.
(a) *Weight and* ***(b)*** *temperature measures of R6/2 mice injected at 4 weeks of age with 3µL of 3.53 x 1013 GC/mL AAV2/9.CB7.eGFP-miR.mFan1 into the left striatum and 5 µL IP (n=9) or 3 µL of 3.78 x 1013 GC/mL AAV2/9.CB7.CI.eGFP-miR.Control into the left striatum and 5 µL IP (n=8). Mice were culled at 7 weeks post transfection. Error bars represent SEM.*

7.5.4.2 Fan1 qPCR

7.5.4.2.1 Liver

The following comparisons were made by loading the indicated samples in random order in triplicate on the same 96 well plate.

1. **8 week samples**. Liver from 8 week R6/2 scrambled or *Fan1* knockdown (n = 11 and 13, respectively)
2. **11 week samples**. Liver from 11 week R6/2 scrambled (n = 10), *Fan1* knockdown (n = 10), or untreated R6/2 (n = 13), or 11 week wild type (WT, n = 12).

Atp5b, *Sdha* and *Ubc* were selected as housekeeping genes and the comparative Ct method was used to calculate relative *Fan1* knockdown. *Fan1* knockdown at 8 weeks was 21% (se = 8.9%, p = 0.0353) and at 11 weeks was 23% (se = 11.8%, p = 0.0671).

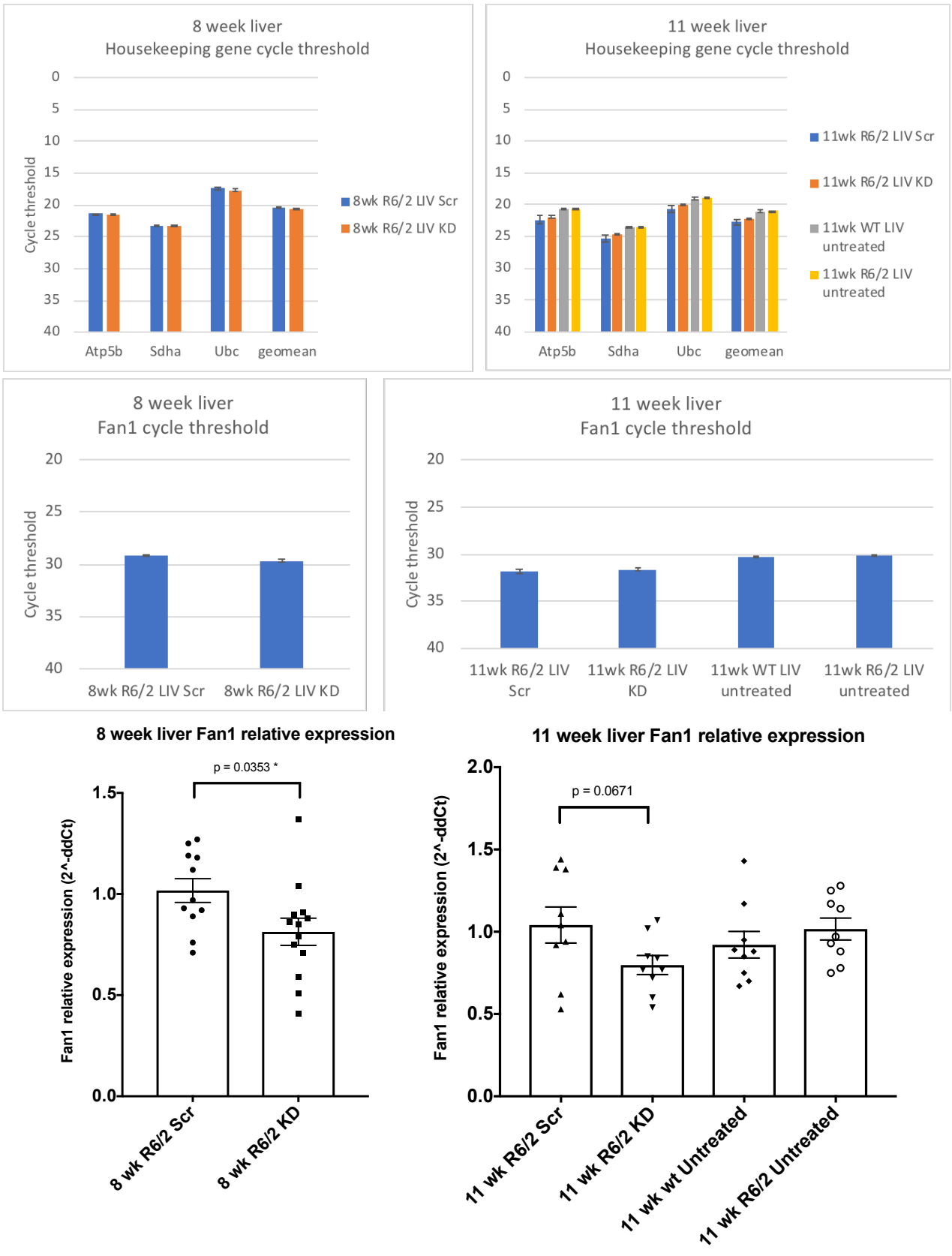


Figure 7.23. Liver Fan1 expression.

Top left *– 8 week liver housekeeping gene cycle threshold,* ***top right*** *– 11 week liver housekeeping gene cycle threshold.* ***Middle left*** *– 8 week liver Fan1 cycle threshold,* ***middle right*** *– 11 week liver Fan1 cycle threshold.* ***Bottom left*** *– 8 week liver Fan1 expression, relative to the mean of 8 week scrambled.* ***Bottom right*** *– 11 week liver Fan1 expression, relative to the mean of 11 week R6/2 untreated. LIV – liver, Scr – scrambled, KD – knockdown.*

7.5.4.2.2 Striatum

The following comparisons were made.

1. **Left vs. right striatum.** A subset of left striatum *scrambled* (n = 6) or *Fan1 knockdown* (n = 8), and right striatum *scrambled* (n = 7) or *Fan1 knockdown* (n = 8) samples. This permits the accurate comparison of left and right striatum, as samples are analysed on the same plates. *Fan1* expression in 8 week samples was calculated relative to 8 week *scrambled* left striatum, and expression in 11 week samples was calculated relative to 11 week *untreated* left striatum.
2. ***Fan1* knockdown in left striatum.** All left striatal samples. 8 week R6/2 *scrambled* (n = 8) or *Fan1 knockdown* (n = 9), 11 week R6/2 *scrambled* (n = 9), *Fan1 knockdown* (n = 9), and untreated (n = 9), and 11 week *untreated* WT (n = 9). *Fan1* expression calculated relative to 11 week *untreated* R6/2.
3. ***Fan1* knockdown in right striatum.** All right striatal samples. 8 week R6/2 *scrambled* (n = 9) or *Fan1 knockdown* (n = 9), 11 week R6/2 *scrambled* (n = 9), *Fan1 knockdown* (n = 9), or untreated (n = 9), and 11 week *untreated* WT (n = 9). Fan1 expression calculated relative to 11 week *untreated* R6/2.

7.5.4.2.2.1 *Fan1* expression in left vs right striatum

To accurately compare *Fan1* expression in the injected left striatum with the uninjected right striatum, a subset of left and right striatum samples, injected with either the scrambled or *Fan1* knockdown virus, were loaded on the same plate.

There was no significant difference between *Fan1* expression level in 11 week **untreated** left and right striatum (mean relative expression 1.024 and 1.063, p = 0.7310), or between left and right striatum treated with the **scrambled** virus at 8 weeks (1.06 and 0.703, p = 0.0791) or 11 weeks (0.919 and 0.845, p = 0.324).

The **active** virus reduced *Fan1* expression in the left striatum at 8 weeks by 29% (p = 0.0513) and at 11 weeks by 25% (p = 0.0506). Later analyses, which focus on left or right striatum and include more animals, improve the power for evaluation of *Fan1* knockdown.

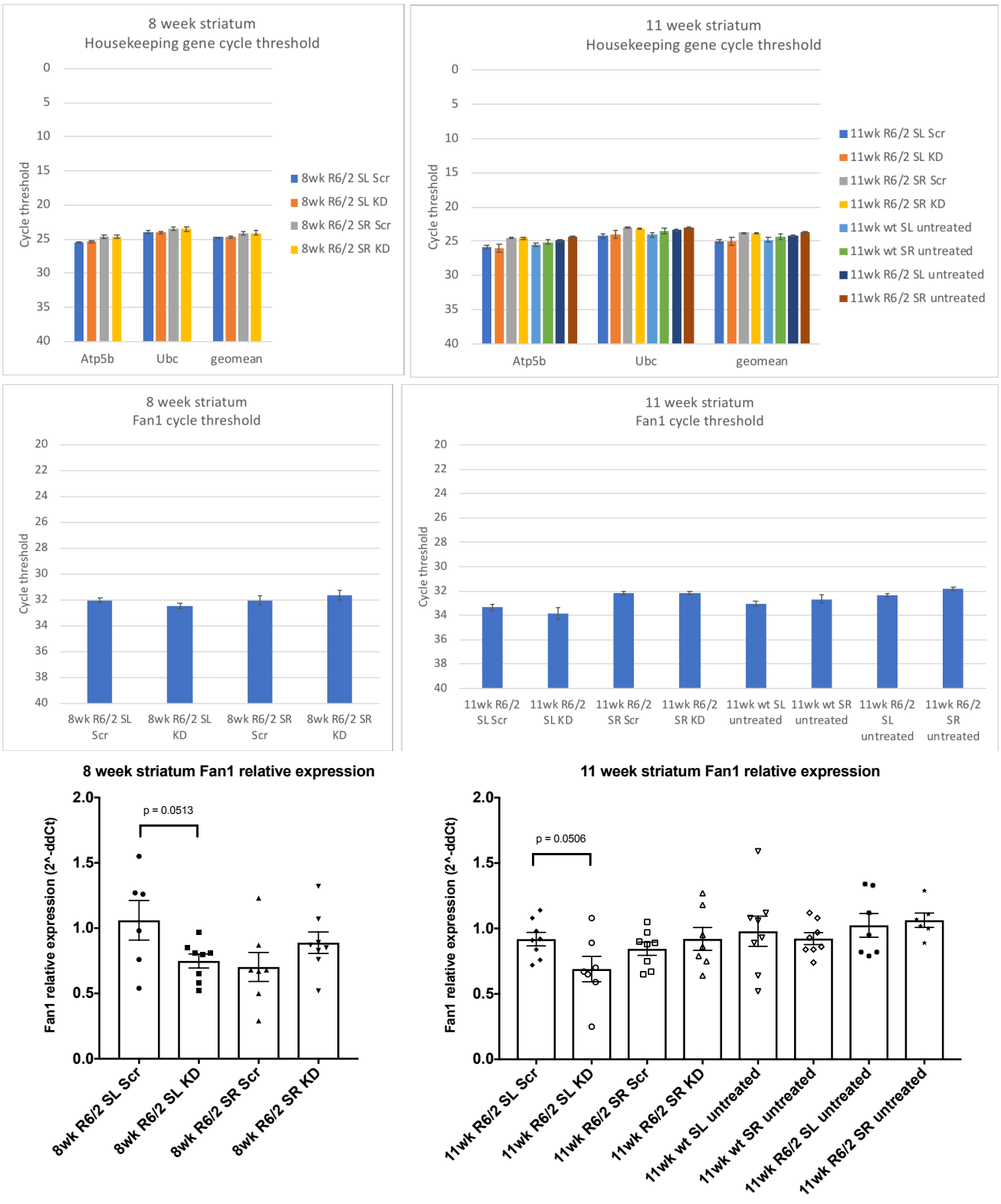


Figure 7.24. Striatum Fan1 relative expression.

Top left *– 8 week striatum housekeeping gene cycle threshold,* ***top right*** *– 11 week striatum housekeeping gene cycle threshold.* ***Middle left*** *8 week striatum Fan1 cycle threshold,* ***middle right*** *– 11 week striatum Fan1 cycle threshold.* ***Bottom left*** *– 8 week striatum Fan1 expression, relative to the mean of 8 week scrambled,* ***Bottom right*** *– 11 week striatum Fan1 expression, relative to the mean of 11 week R6/2 untreated. SL – left striatum, SR – right striatum, Scr – scrambled, KD – knockdown.*

7.5.4.2.2.2 *Fan1* knockdown in left and right striatum

To accurately quantify *Fan1* knockdown, all samples from left and right striatum were analysed together. The active virus reduced *Fan1* expression in 8 week **left striatum** by 19% (se = 9.0%, p = 0.0209) and in 11 week left striatum by 28% (se = 7.8%, p = 0.0024). There was no significant knockdown in 8 or 11 week **right striatum** (p = 0.3658 and 0.8405), and once again *Fan1* expression in the 8 and 11 week **scrambled** treated right striatum did not significantly differ from 11 week untreated R6/2.

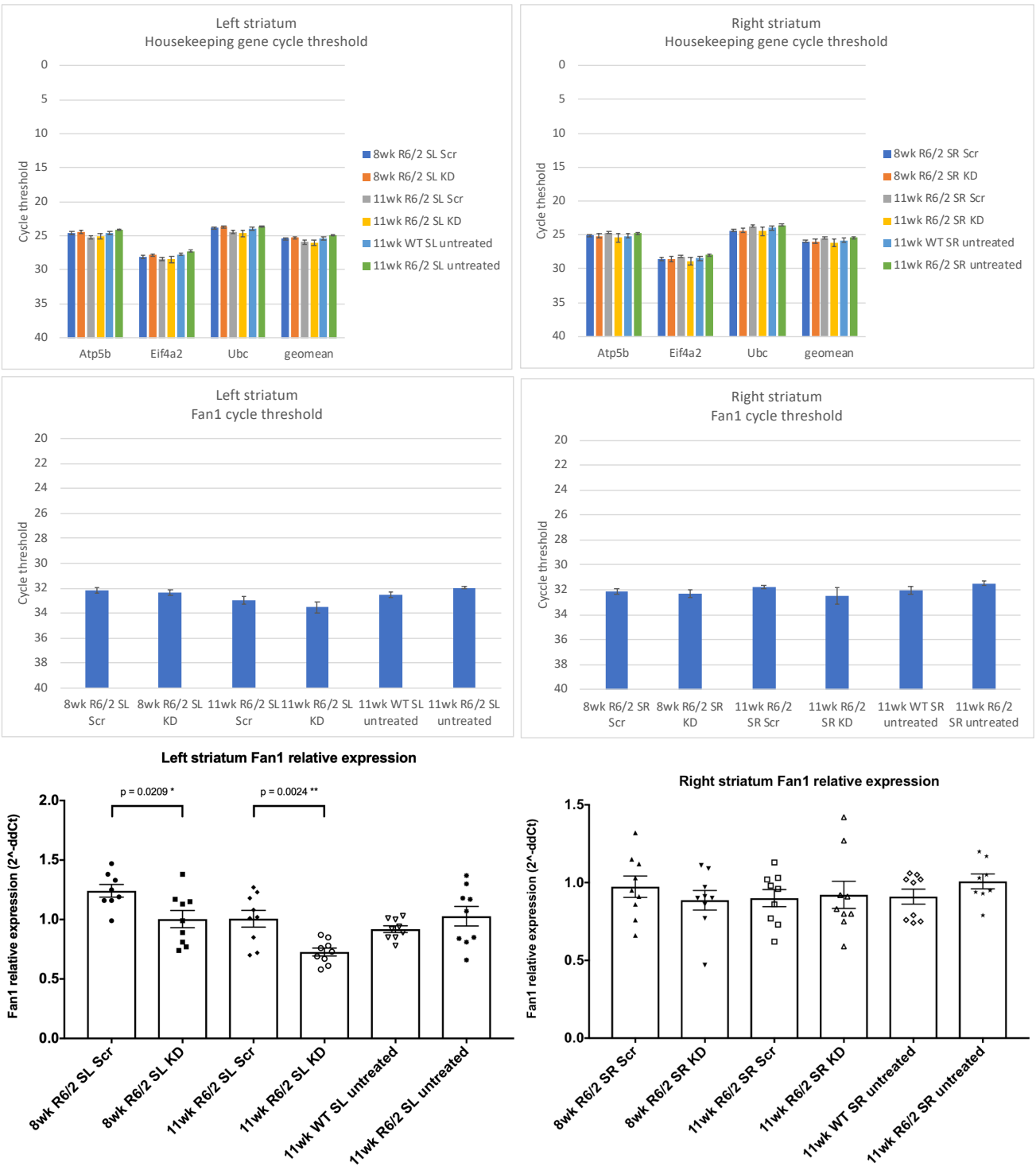


Figure 7.25. Striatum Fan1 expression.

Top left *– left striatum housekeeping gene cycle threshold,* ***top right*** *– right striatum housekeeping gene cycle threshold.* ***Middle left*** *– left striatum Fan1 cycle threshold,* ***middle right*** *– right striatum Fan1 cycle threshold.* ***Bottom left*** *– left striatum Fan1 expression, relative to 11 week untreated R6/2 left striatum.* ***Bottom right*** *– right striatum Fan1 expression, relative to 11 week untreated R6/2 right striatum. SL – left striatum, SR – right striatum, Scr – scrambled, KD – knockdown.*

7.5.4.2.3 Conclusions

The active virus reduced *Fan1* expression on average in left striatum and liver throughout the experiment by 22.7% (sem = 1.9%). Expression in right striatum was unaffected.

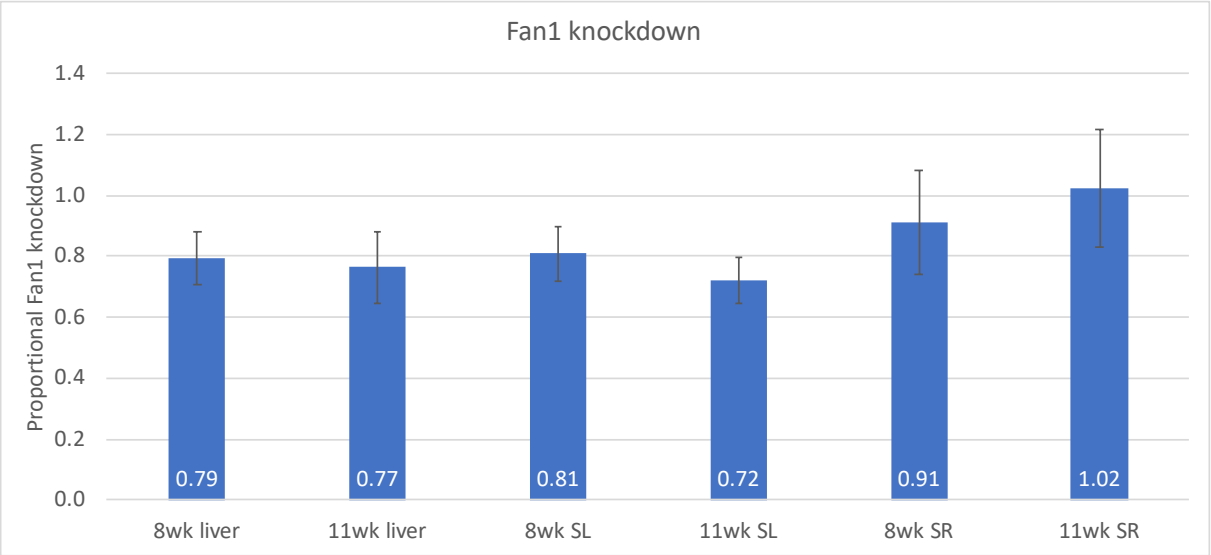


Figure 7.26. Fan1 knockdown in R6/2 tissues.
Fan1 KD expressed as a proportion of scrambled ± SE. SL – striatum left, SR – striatum right.

7.5.4.3 CAG repeat sizing

7.5.4.3.1 Modal CAG repeat length

CAG repeat size was determined by fragment analysis using the Bates lab protocol (see Methods). Each sample was sized three times independently, then averaged. In 12 day tail at baseline, there was no significant difference between the CAG repeat size in *Fan1* knockdown and scrambled ($p = 0.721$) or untreated animals ($p = 0.637$). Mean repeat sizes were 188.0 (95% CI 185.7-190.3), 188.4 (95% CI 187.5-189.4) and 187.8 (95% CI 183.9-191.7) respectively.

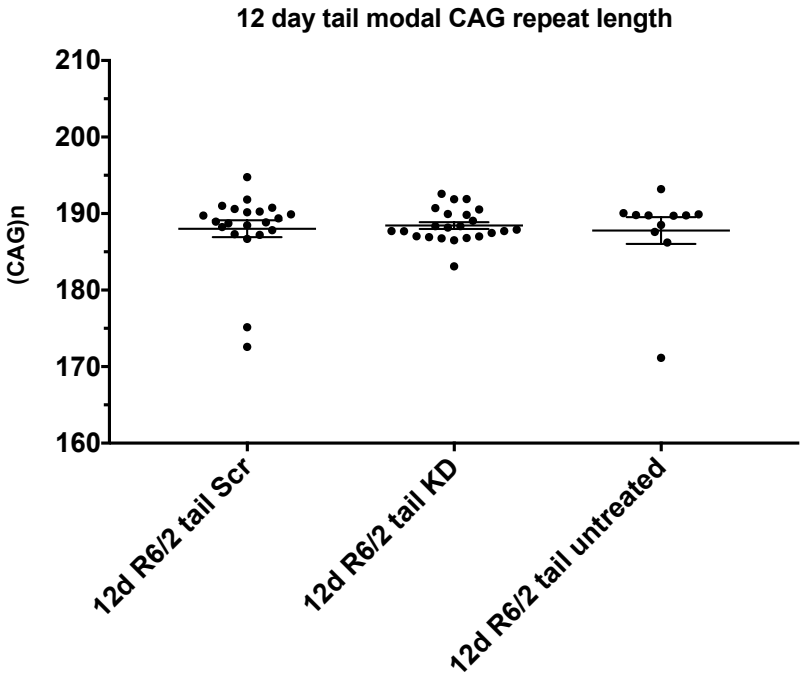


Figure 7.27. Modal CAG repeat size at baseline (12 day tail).
Error bars represent SEM. Scr – scrambled, KD – knockdown.

Each tissue is considered in turn below.

7.5.4.3.1.1 Tail

Modal CAG repeat length was not significantly different between 11 week scrambled and untreated tail, suggesting viral transduction itself did not influence CAG repeat length (p = 0.236). CAG length increased in scrambled 8 week tail on average by 0.83 repeats (sem = 0.40, p = 7.13E-3), in 11 week scrambled tail by 0.97 (sem = 0.39, p = 5.01E-4) and in 11 week untreated tail by 0.36 (sem = 0.26, p = 5.74E-2). Modal change was not significantly different between scrambled and *Fan1* knockdown 8 week tail (p = 0.636) or 11 week tail (p = 0.637), suggesting *Fan1* knockdown did not affect expansion in the tail.

7.5.4.3.1.2 Left striatum

For left striatum, there was no significant difference between 11 week scrambled and untreated animals (p = 0.236), again showing the virus itself had no effect on expansion. The CAG increased in scrambled 8 week left striatum on average by 1.79 repeats (sem = 0.22, p = 3.97E-12), in 11 week scrambled left striatum by 1.81 repeats (sem = 0.33, p = 9.43E-9) and in 11 week untreated left striatum by 0.91 (sem = 0.50, p = 1.64E-2). There was no significant difference in modal change between scrambled and *Fan1* knockdown 8 week left striatum (p = 0.636), or 11 week left striatum (p = 0.706), suggesting *Fan1* knockdown did not affect expansion in the left striatum.

7.5.4.3.1.3 Right striatum

For right striatum, there was no significant difference between 11 week scrambled and untreated animals (p = 0.757). The CAG increased in scrambled 8 week right striatum on average by 1.88 repeats (sem = 0.35, p = 2.43E-8), in 11 week scrambled right striatum by 2.10 repeats (sem = 0.47, p = 2.91E-7) and in 11 week untreated right striatum by 1.87 (sem = 0.50, p = 1.21E-5). There was no significant difference in modal change between scrambled and *Fan1* knockdown 8

week right striatum (p = 0.6073), or 11 week right striatum (p = 0.424), suggesting *Fan1* knockdown did not affect expansion in the right striatum.

7.5.4.3.1.4 Liver

In liver, there was no significant difference between 11 week scrambled and untreated animals (p = 0.532). The CAG increased in scrambled 8 week liver on average by 2.66 repeats (sem = 0.5, p = 3.35E-8), in 11 week scrambled liver by 3.29 repeats (sem = 1.24, p = 1.05E-3) and in 11 week untreated liver by 1.97 (sem = 1.38, p = 0.053). There was no significant difference in modal change between scrambled and *Fan1* knockdown 8 week liver (p = 0.708), or 11 week liver (p = 0.314), suggesting *Fan1* knockdown did not affect expansion in the liver.

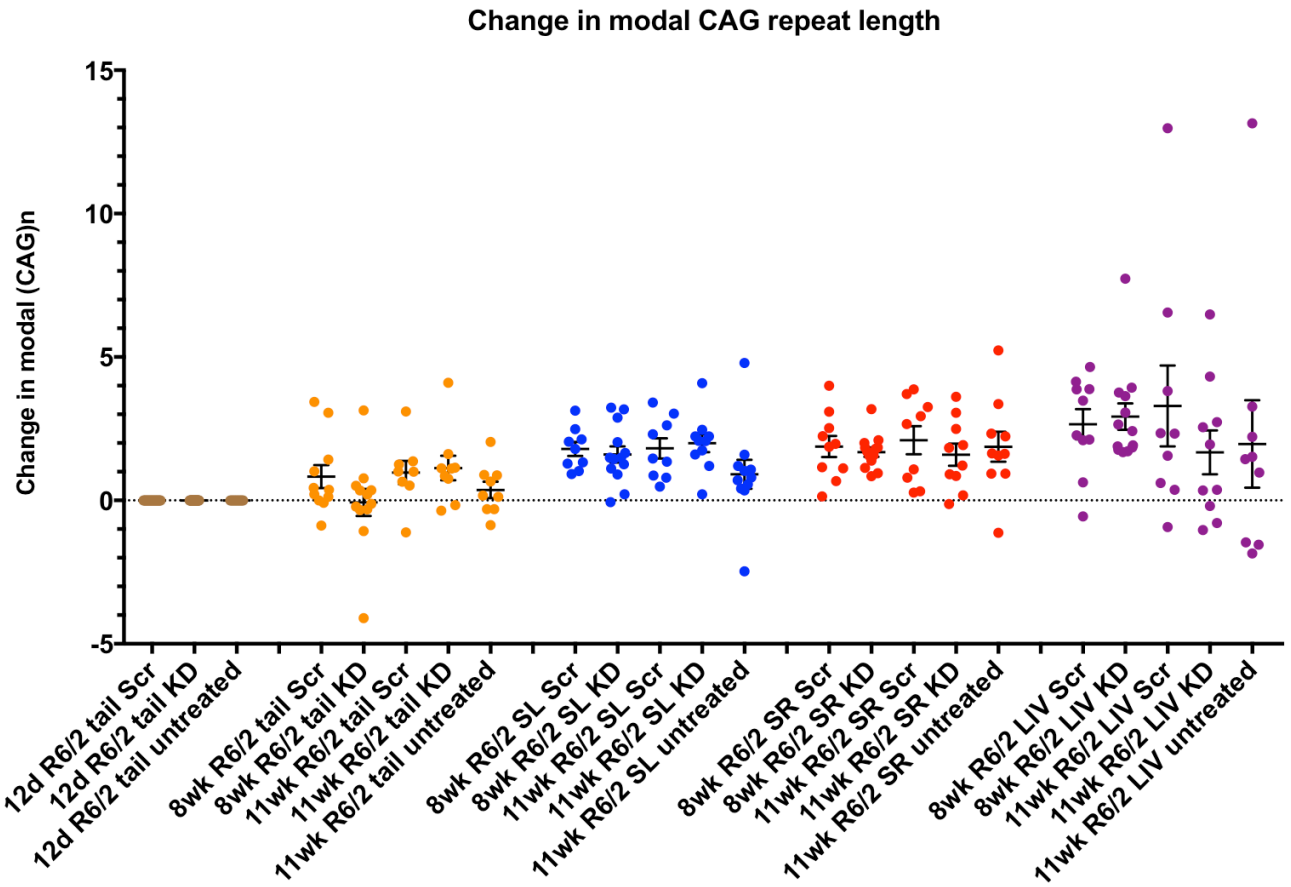


Figure 7.28. Change in modal CAG repeat length relative to 12 day tail.
Change in modal CAG repeat length is calculated relative to each animal's own 12d tail. Error bars represent SEM. Brown – 12 day tail, orange – 8-11 week tail, blue – left striatum (SL), red – right striatum (SR), purple – liver (LIV). Scr – scrambled. KD – knockdown.

Treatment	n	Modal (CAG)n	SEM (mode)	Change in modal (CAG)n	SEM (Δmode)
12d R6/2 tail Scr	21	188.02	1.10	0.00	0.00
12d R6/2 tail KD	23	188.43	0.45	0.00	0.00
12d R6/2 tail untreated	11	187.79	1.74	0.00	0.00
8wk R6/2 tail Scr	11	190.12	0.68	0.83	0.40
8wk R6/2 tail KD	13	188.64	0.74	-0.07	0.45
11wk R6/2 tail Scr	9	188.90	1.85	0.97	0.39
11wk R6/2 tail KD	9	189.36	0.47	1.13	0.43
11wk R6/2 tail untreated	11	187.71	1.95	0.36	0.26
8wk R6/2 SL Scr	11	191.01	0.74	1.79	0.22
8wk R6/2 SL KD	13	190.18	0.79	1.60	0.28
11wk R6/2 SL Scr	10	188.04	2.29	1.81	0.33
11wk R6/2 SL KD	10	190.23	0.45	1.99	0.31
11wk R6/2 SL untreated	11	188.70	1.88	0.91	0.50
8wk R6/2 SR Scr	11	191.43	0.76	1.88	0.35
8wk R6/2 SR KD	13	190.27	0.80	1.68	0.16
11wk R6/2 SR Scr	10	188.33	2.27	2.10	0.47
11wk R6/2 SR KD	10	189.83	0.46	1.59	0.38
11wk R6/2 SR untreated	11	189.46	1.97	1.87	0.50
8wk R6/2 LIV Scr	11	192.21	0.71	2.66	0.50
8wk R6/2 LIV KD	13	191.50	0.93	2.92	0.46
11wk R6/2 LIV Scr	10	189.52	1.40	3.29	1.34
11wk R6/2 LIV KD	10	189.91	0.84	1.67	0.76
11wk R6/2 LIV untreated	11	191.38	1.40	1.97	1.38

Table 7.4. Change in modal CAG relative to 12 day tail.

SL – left striatum, SR – right striatum, LIV – liver, Scr – scrambled, KD – knockdown, dmode – change in modal CAG repeat length.

7.5.4.3.2 Somatic instability index

The somatic instability index (SII), which is measured in CAG repeat units, is considered a more sensitive measure of change in repeat length (Lee et al., 2010, Zhao and Usdin, 2018).

7.5.4.3.2.1 Scrambled and untreated tissues

SII increased in scrambled 8 week tail by 1.30 (sem = 0.45, p = 3.34E-4), in 11 week scrambled tail by 1.82 (sem = 0.48, p = 9.44E-7), in 11 week untreated tail by 1.17 (sem = 0.20, p = 2.39E-6), in 8 week scrambled left striatum by 3.50 (sem = 0.47, p = 2.45E-12), in 11 week scrambled left striatum by 7.46 (sem = 0.93, p = 4.68E-13), in 11 week untreated left striatum by 6.16 (sem = 0.91, p = 1.58E-10), in 8 week scrambled right striatum by 6.27 (sem 0.38, p = <1.00E-15), in 11 week scrambled week right striatum by 9.03 (sem = 1.10, p = 3.15E-13), in 11 week untreated right striatum by 5.867 (sem = 1.02, p = 2.37E-9), in 8 week scrambled liver by 4.40 (sem = 0.31, p = <1.00E-15), in 11 week scrambled liver by 6.56 (sem = 0.44, p = <1.00E-15) and in 11 week untreated liver by 6.84 (sem = 0.79, p = 8.40E-14).

There was significantly more expansion in scrambled left striatum compared to tail at 8 weeks (p = 3.47E-3) and 11 weeks (p = 1.08E-4). Untreated left striatum expanded significantly more than tail at 11 weeks (p = 1.60E-4). There was significantly more expansion in right compared to left scrambled striatum at 8 weeks (p = 2.52E-4), but not at 11 weeks (p = 0.29). Expansion was not significantly different between scrambled left striatum and liver at 8 weeks (p = 0.12) or 11 weeks (p = 0.39), or between untreated left striatum and liver at 11 weeks (p = 0.59).

7.5.4.3.2.2 *Fan1* knockdown

Comparing scrambled and *Fan1* knockdown, there was no significant difference in tail at 8 weeks (p = 0.49) or 11 weeks (p = 0.50), in left striatum at 8 weeks (p = 0.10) or 11 weeks (p = 0.28), in right striatum at 8 weeks (p = 0.26) or 11 weeks (p = 0.34), or in liver at 8 weeks (p = 0.86) or 11 weeks (p = 0.32).

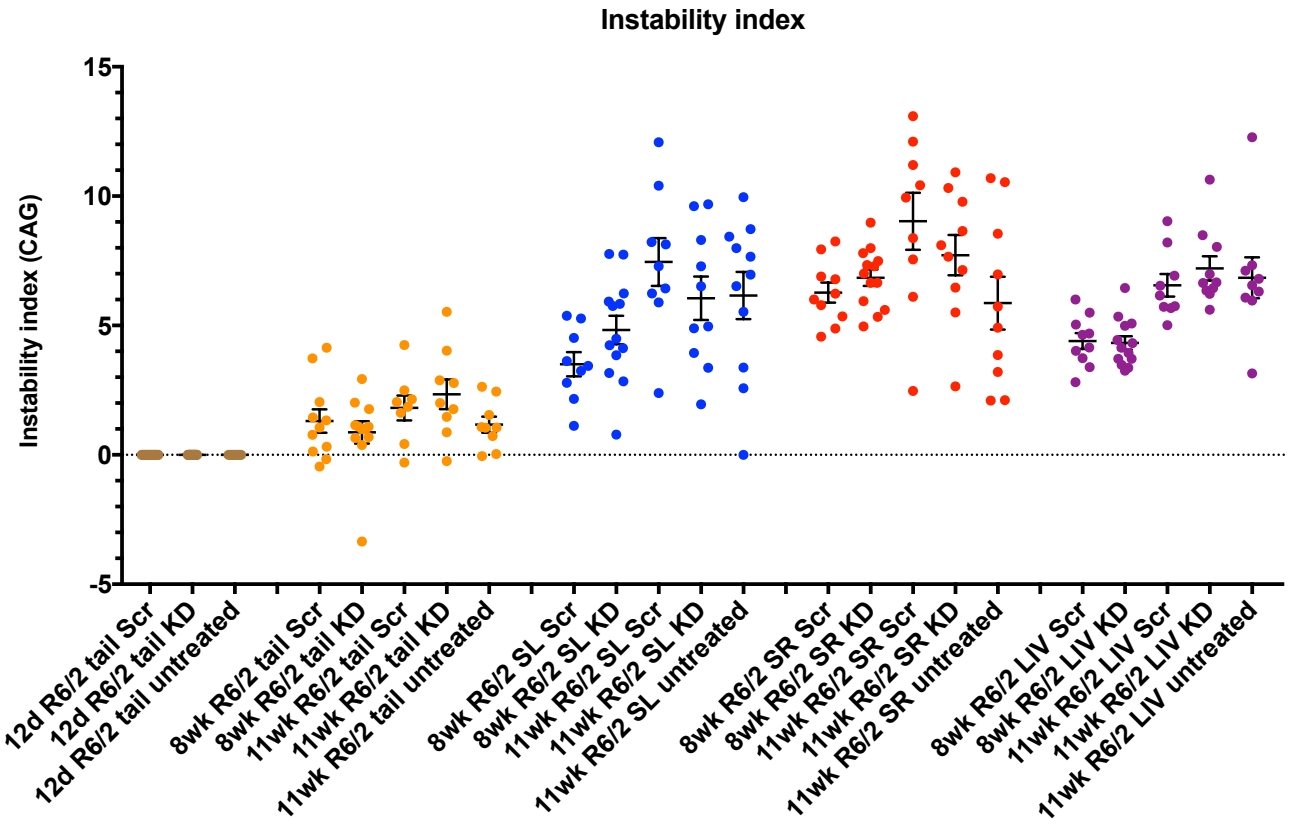


Figure 7.29. Somatic instability index relative to 12 day tail.
Instability index is calculated relative to each animal's own 12d tail. Error bars represent SEM. Brown – 12 day tail, orange – 8-11 week tail, blue – left striatum (SL), red – right striatum (SR), purple – liver (LIV). Scr – scrambled. KD – knockdown.

Treatment	n	Somatic Instability index (SII)	SEM (sii)
12d R6/2 tail Scr	21	0.00	0.00
12d R6/2 tail KD	23	0.00	0.00
12d R6/2 tail untreated	11	0.00	0.00
8wk R6/2 tail Scr	11	1.30	0.45
8wk R6/2 tail KD	13	0.87	0.42
11wk R6/2 tail Scr	9	1.82	0.45
11wk R6/2 tail KD	9	2.34	0.57
11wk R6/2 tail untreated	11	1.17	0.28
8wk R6/2 SL Scr	11	3.50	0.42
8wk R6/2 SL KD	13	4.83	0.55
11wk R6/2 SL Scr	10	7.45	0.88
11wk R6/2 SL KD	10	6.05	0.84
11wk R6/2 SL untreated	11	6.16	0.91
8wk R6/2 SR Scr	11	6.27	0.37
8wk R6/2 SR KD	13	6.85	0.32
11wk R6/2 SR Scr	10	9.03	1.05
11wk R6/2 SR KD	10	7.72	0.78
11wk R6/2 SR untreated	11	5.87	0.97
8wk R6/2 LIV Scr	11	4.40	0.29
8wk R6/2 LIV KD	13	4.32	0.26
11wk R6/2 LIV Scr	10	6.56	0.41
11wk R6/2 LIV KD	10	7.21	0.47
11wk R6/2 LIV untreated	11	6.84	0.72

Table 7.5. Somatic instability index relative to 12d tail.
SL – left striatum, SR – right striatum, LIV – liver, Scr – scrambled, KD – knockdown, sii – somatic instability index.

7.5.4.3.3 Proportional expansion analysis

7.5.4.3.3.1 Scrambled and untreated tissues

The proportion of the distribution width greater than the baseline mode (12d tail) increased significantly in 8 week scrambled tail (p = 1.02E-4), 11 week scrambled tail (p = 1.68E-6), 11 week untreated tail (p = 1.74E-3), 8 week scrambled left striatum (p = 9.37E-8), 11 week scrambled left striatum (p = 1.49E-11), 11 week untreated left striatum (p = 4.85E-8), 8 week scrambled right striatum (p = 1.10E-14), 11 week scrambled right striatum (p = 3.40E-13), 11 week untreated right striatum (p = 1.01E-5), 8 week scrambled liver (p = 1.14E-11), 11 week scrambled liver (p = 4.50E-14) and 11 week untreated liver (p = 3.27E-6).

Scrambled left striatum expanded significantly more than tail at 8 weeks (p = 0.049) and 11 weeks (p = 0.012). Untreated left striatum expanded significantly more than tail at 11 weeks (p = 8.22E-5). There was no significant difference between scrambled left and right striatum at 8 weeks (p = 0.21) or 11 weeks (p = 0.21), or between untreated left and right striatum at 11 weeks (p = 0.46). There was no significant difference between scrambled left striatum and liver at 8 weeks (p = 0.97) or 11 weeks (p = 0.86), or between untreated left striatum and liver at 11 weeks (p = 0.29).

7.5.4.3.3.2 *Fan1* knockdown

Comparing scrambled and *Fan1* knockdown animals, there was no significant difference in tail at 8 weeks (p = 0.46) or 11 weeks (p = 0.85), in left striatum at 8 weeks (p = 0.77) or 11 weeks (p = 0.67), in right striatum at 8 weeks (p = 0.28) or 11 weeks (p = 0.51), or in liver at 8 weeks (p = 0.50) or 11 weeks (p = 0.77).

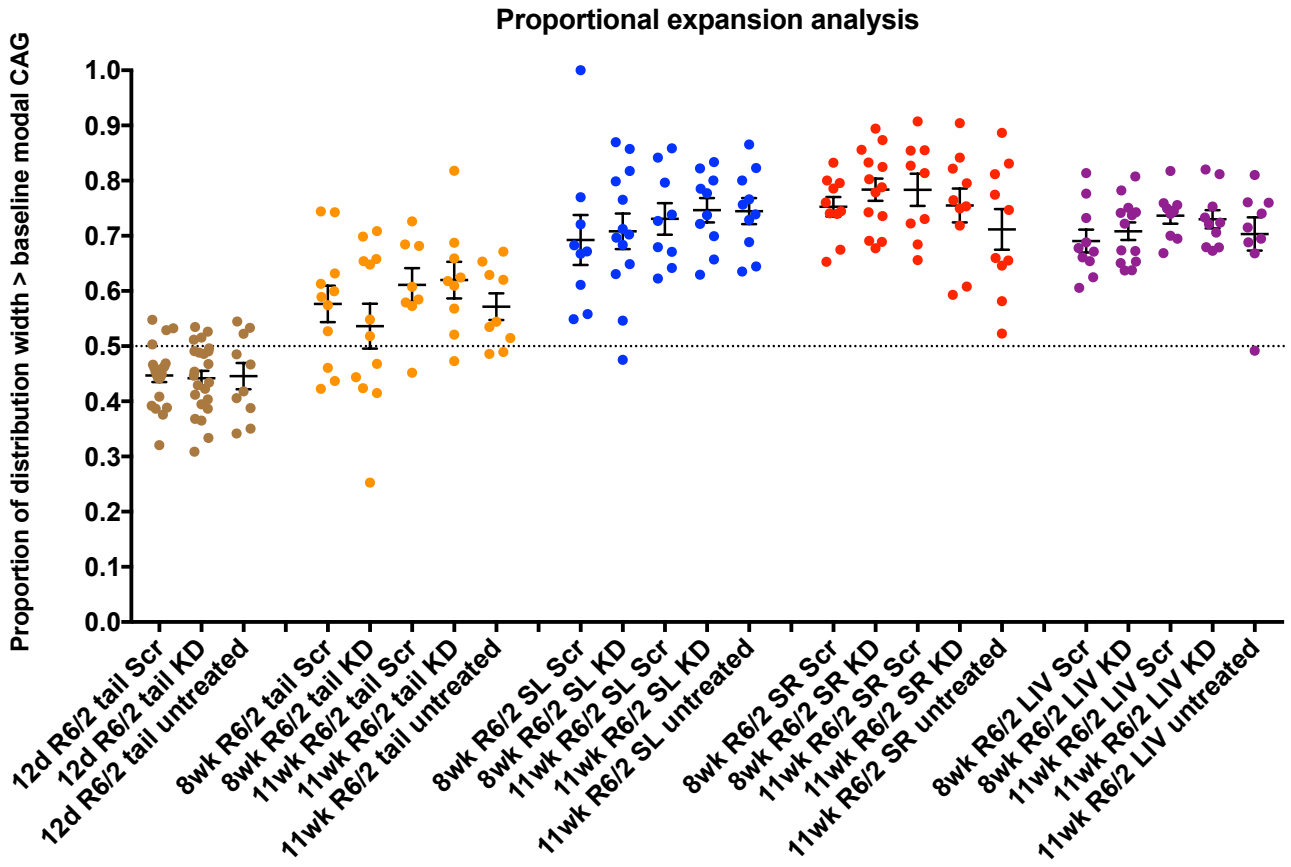


Figure 7.30. Proportional expansion analysis.

The proportion of the distribution width greater than the baseline (12d tail) modal CAG length. 0.5 represents a normal distribution with a mode equal to the control mode. The maximum is 1.0 (the entire distribution is greater than the control mode) and minimum is 0.0 (the entire distribution is less than the control mode). Error bars represent SEM. Brown – 12 day tail, orange – 8-11 week tail, blue – left striatum (SL), red – right striatum (SR), purple – liver (LIV). Scr – scrambled. KD – knockdown.

Treatment	n	Proportion of distribution > control mode	SEM (proportion)
12d R6/2 tail Scr	21	0.45	0.01
12d R6/2 tail KD	23	0.44	0.01
12d R6/2 tail untreated	11	0.45	0.02
8wk R6/2 tail Scr	11	0.58	0.03
8wk R6/2 tail KD	13	0.54	0.04
11wk R6/2 tail Scr	9	0.61	0.03
11wk R6/2 tail KD	9	0.62	0.03
11wk R6/2 tail untreated	11	0.57	0.02
8wk R6/2 SL Scr	11	0.69	0.04
8wk R6/2 SL KD	13	0.71	0.03
11wk R6/2 SL Scr	10	0.73	0.03
11wk R6/2 SL KD	10	0.75	0.02
11wk R6/2 SL untreated	11	0.74	0.02
8wk R6/2 SR Scr	11	0.75	0.02
8wk R6/2 SR KD	13	0.78	0.02
11wk R6/2 SR Scr	10	0.78	0.03
11wk R6/2 SR KD	10	0.76	0.03
11wk R6/2 SR untreated	11	0.71	0.04
8wk R6/2 LIV Scr	11	0.69	0.02
8wk R6/2 LIV KD	13	0.71	0.02
11wk R6/2 LIV Scr	10	0.74	0.01
11wk R6/2 LIV KD	10	0.73	0.02
11wk R6/2 LIV untreated	11	0.70	0.03

Table 7.6. Proportional expansion analysis.

The proportion of the distribution width greater than the baseline (12d tail) modal CAG length. 0.5 represents a normal distribution with a mode equal to the control mode. The maximum is 1.0 (the entire distribution is greater than the control mode) and minimum is 0.0 (the entire distribution is less than the control mode). SL – left striatum, SR – right striatum, LIV – liver, Scr – scrambled. KD – knockdown.

7.5.4.4 Comparing CAG repeat expansion and Fan1 expression

7.5.4.4.1 Change in modal CAG repeat length

Change in modal CAG repeat length was regressed against *Fan1* expression level. No significant effect was observed in left striatum, right striatum or liver. Grouping liver tissue at all ages, there was a trend towards reduced expansion with higher *Fan1* expression (slope = -2.54 $\pm$ 1.48, p = 0.093).

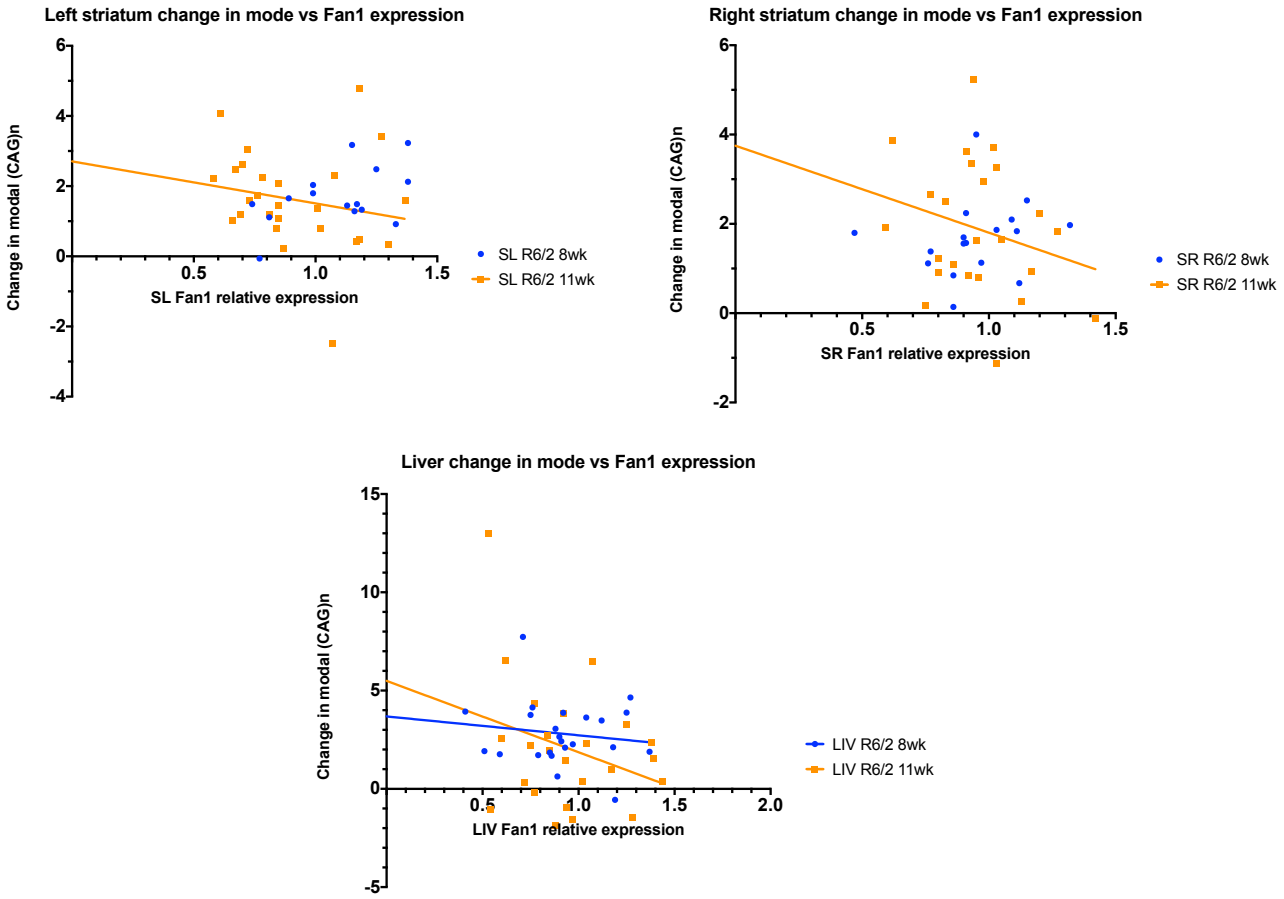


Figure 7.31. Change in modal CAG repeat length against Fan1 expression.
Top left *– left striatum (SL),* ***top right*** *– right striatum (SR),* ***bottom*** *– liver (LIV). Lines represent linear regression. Blue – 8 week R6/2, orange – 11 week R6/2.*

7.5.4.4.2 Somatic instability index

There was no significant correlation between somatic instability index (SII) and *Fan1* expression level.

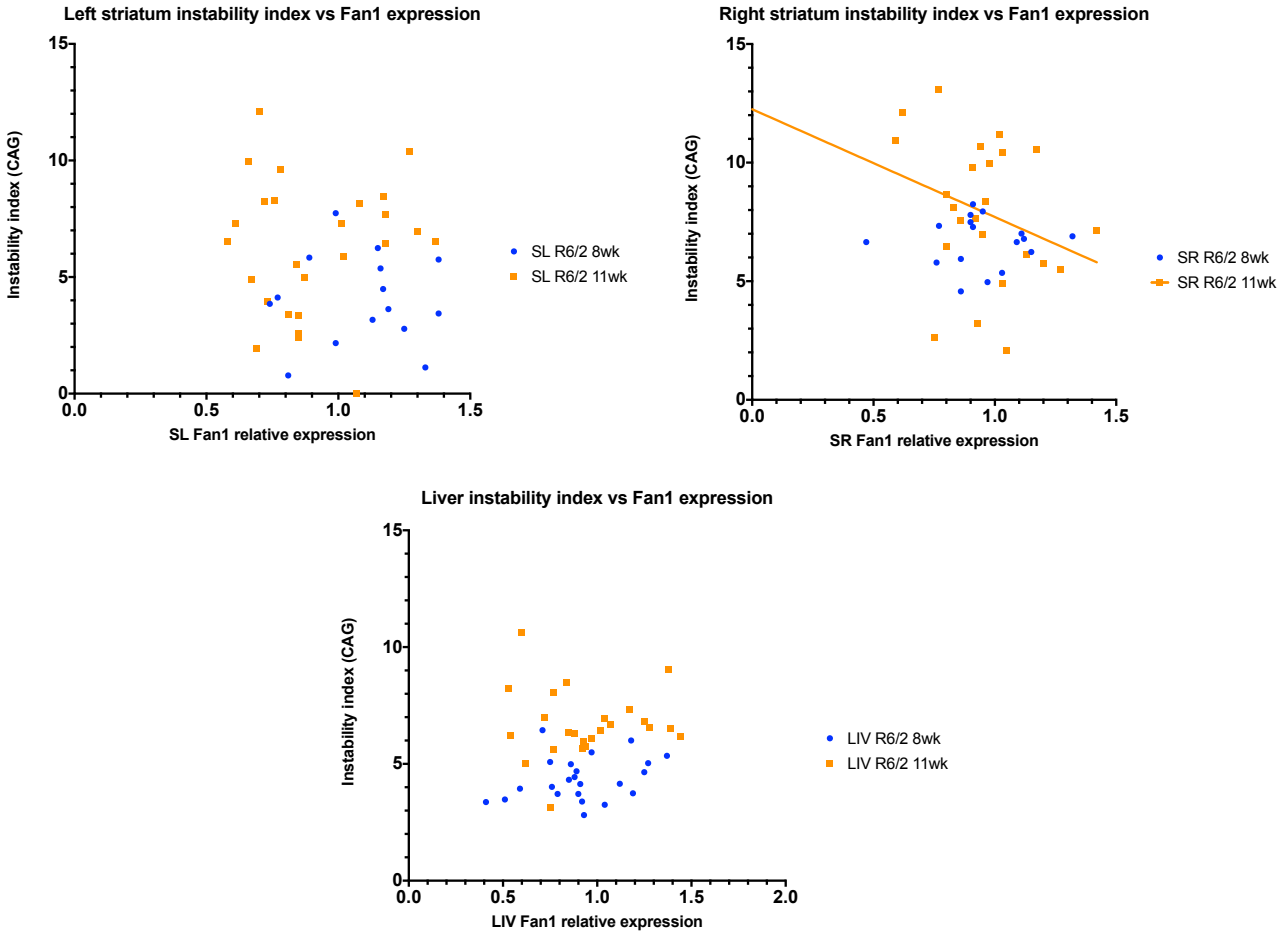


Figure 7.32. Somatic instability index against Fan1 expression.
Top left *– left striatum (SL),* ***top right*** *– right striatum (SR),* ***bottom*** *– liver (LIV). Lines represent linear regression. Blue – 8 week R6/2, orange – 11 week R6/2.*

7.5.4.4.3 Proportional expansion analysis

The proportional expansion metric was not significantly associated with *Fan1* expression.

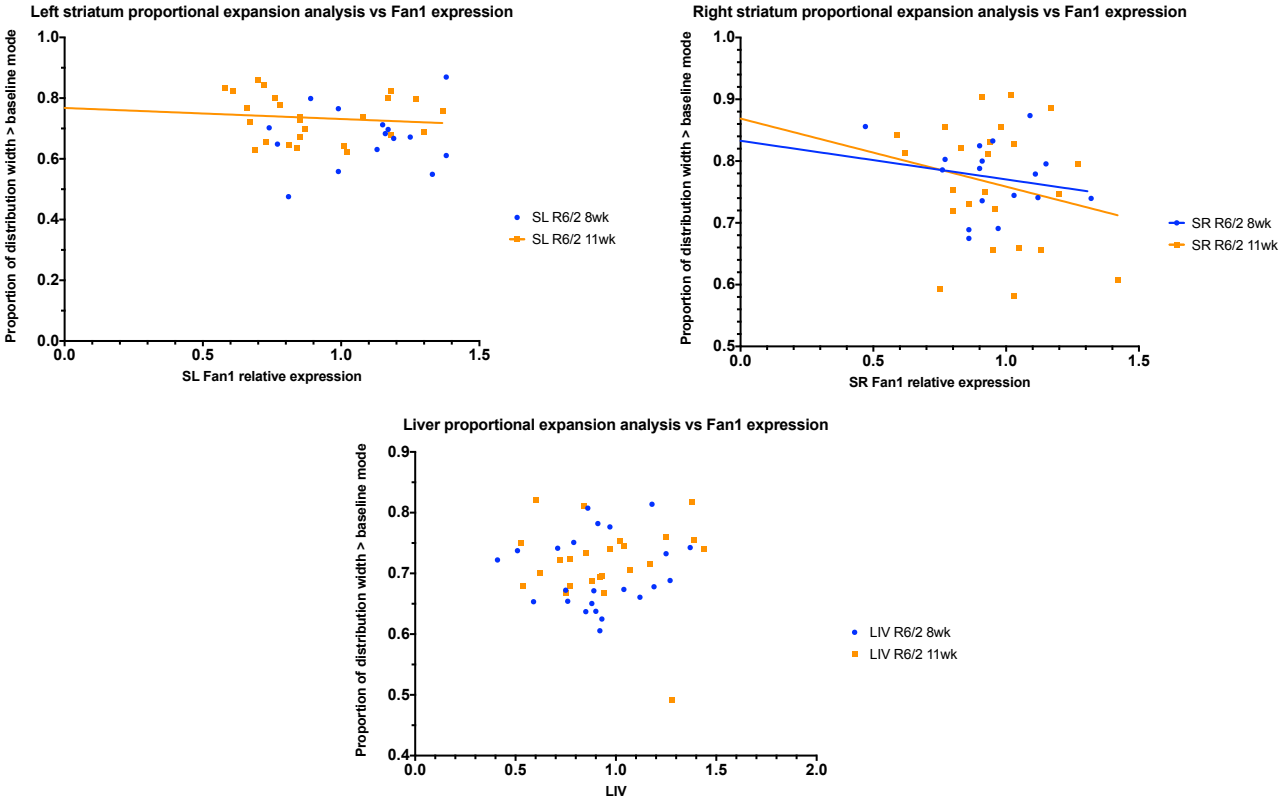


Figure 7.33. Proportional expansion analysis against Fan1 expression.
The proportion of the distribution width greater than the baseline (12d tail) modal CAG length. 0.5 represents a normal distribution with a mode equal to the control mode. The maximum is 1.0 (the entire distribution is greater than the control mode) and minimum is 0.0 (the entire distribution is less than the control mode). ***Top left*** *– left striatum (SL),* ***top right*** *– right striatum (SR),* ***bottom*** *– liver (LIV). Lines represent linear regression. Blue – 8 week R6/2, orange – 11 week R6/2.*

7.6 Discussion

7.6.1 Fan1 expression level in R6/2 tissues

Our pilot study of *Fan1* expression in ten R6/2 tissues at 4 and 14 weeks showed relatively high transcript levels in cerebellum, cord, hippocampus and lung, and relatively low levels in adrenal, kidney and liver, all of which are consistent with published microarray and RNA-Seq data (EMBL-EBI Expression Atlas) (Papatheodorou et al., 2018). CAG expansion is known to be marked in the liver and limited in the cerebellum, suggesting *Fan1* expression may be a protective modifier.

7.6.2 Optimisation of assays

This project required DNA, RNA and protein extraction from a small 9 mg striatum sample. A 3-in-1 extraction kit was trialled, which uses filter column technology to sequentially extract DNA, RNA and protein from the same sample. Though DNA and RNA purity were equivalent to traditional extraction techniques, the yield was significantly reduced to 4% and 28% respectively. *Fan1* relative expression, as calculated by the comparative Ct method, was significantly different between 3-in-1 and traditionally extracted cDNA. The protein yield was also significantly reduced, and nuclear proteins were selectively lost, possibly because nuclear fractions were removed with DNA and RNA during the 3-in-1 protocol.

Traditional DNA, RNA and protein extraction, followed by CAG repeat sizing, RT-qPCR and western blot demonstrated that robust data can be generated from 1/3 of a striatum. Therefore, experimental striatal samples were divided into thirds and extracted using traditional methods.

7.6.3 Toxicity study

The left striatum was successfully transduced by direct injection, as demonstrated on western blot and immunohistochemistry by high GFP expression and *Fan1* knockdown by around 90%. Though not directly injected, the right striatum was transduced at a lower level and showed around 50% *Fan1* knockdown. *Fan1* expression levels in the liver are endogenously low, making interpretation of knockdown difficult (Papatheodorou et al., 2018). Following intravenous administration there was modest GFP expression in liver, and lower level expression following striatal injection. Intraperitoneal injection effectively transduced the liver, as demonstrated by immunohistochemistry, though GFP level was lower in the scrambled virus. Neither striatal or intravenous method produced toxicity. There was no difference in HTT aggregate levels between scrambled or *Fan1* knockdown animals.

Given the limited transduction efficiency of intravenous injection, a combined striatal and intraperitoneal injection protocol was selected for the experimental study, aiming to maximise knockdown *Fan1* in the striatum and liver; the tissues most prone to CAG repeat expansion in R6/2.

7.6.4 *Fan1* knockdown

Mice were treated by left striatal and intraperitoneal injection at 4 weeks age which produced stable 23% transcript knockdown, relative to scrambled control, in left striatum and liver throughout the experiment. *Fan1* expression in right striatum was unaffected. This is a significantly smaller effect than observed in the toxicity experiment.

7.6.5 CAG repeat instability

R6/2 mice at the 12 day baseline in each of the scrambled, knockdown and untreated groups had a mean tail CAG repeat length of 188, with no significant difference between groups. Three measures of somatic instability were used, as detailed in the Methods chapter.

1. Change in modal repeat size.
2. Somatic instability index (SII).
3. Proportional expansion analysis.

The modal CAG repeat size increased by around 2 repeat units in left striatum (1.81 ± 0.33), right striatum (2.10 ± 0.47) and liver (3.29 ± 1.34), and by 0.97 ± 0.39 units in tail over the course of the experiment. Using this measure, there was significantly more expansion in striatum and liver than in tail (p = 0.026), but there was no significant difference between striatum and liver. Critically, there was no significant difference between *Fan1* knockdown, scrambled or untreated animals at either 8 or 11 weeks age in any of the tissues sampled.

SII is expressed in CAG repeat units and is a more sensitive measure of instability (Lee et al., 2010, Zhao and Usdin, 2018). Once again this showed significantly more expansion in left striatum (7.45 ± 0.88Q), right striatum (9.03 ± 1.05Q) and liver (6.56 ± 0.41Q) than in tail (1.82 ± 0.45Q) at 11 weeks age (p = 4.68E-13). There was no significant difference between expansion rate in liver and striatum. Once again, *Fan1* knockdown did not significantly alter the SII in any of the tissues.

The proportional expansion method measures the width of the CAG repeat distribution above the modal CAG length of the baseline 12d tail sample. It is expressed as a proportion, with 0.5 indicating a normal distribution centred on the baseline modal CAG length. There was significantly more expansion in left striatum (0.73 ± 0.03), right striatum (0.78 ± 0.02) and liver (0.74 ± 0.01) than in tail (0.61 ± 0.03) at 11 weeks (p = 0.012). Again, there was no significant difference between expansion rate in striatum and liver. Concurring with the previous analyses, *Fan1* knockdown did not significantly alter proportional expansion relative to scrambled or untreated animals.

Several lines of evidence, including in patient-derived stem cell and differentiated medium spiny neurons (Chapters 5 and 6), suggest *Fan1* knockdown accelerates CAG repeat expansion *in vitro*. The lack of effect in this mouse model is likely due to the low level of knockdown. By comparison, stable shRNA-mediated 50% *Fan1* knockdown in a stem cell model with ~120 CAG repeats accelerated expansion rate from 13.95 ± 0.31 days/Q to 9.81 ± 0.27 days/Q (p = 3.15E-15).

Regressing CAG expansion against *Fan1* expression in left and right striatum and liver, there was no significant correlation. However, there did appear to be a trend towards slower expansion in animals with higher *Fan1* expression at 11 weeks in all tissues.

7.7 Summary

Fan1 is expressed at relatively low level in some tissues that show marked CAG repeat expansion, such as liver, and at relatively high levels in some tissues in which expansion is limited, such as cerebellum. This warrants a more detailed exploration of *Fan1* expression and CAG expansion across mouse tissue.

Traditional extraction methods for DNA, RNA and give higher and purer yields than a combined extraction kit, with which we found selective loss of transcripts and nuclear proteins.

The toxicity study suggested left striatal injection gave up to 90% *Fan1* knockdown in the left striatum and 50% knockdown in the right striatum. Liver was well transduced following intraperitoneal injection of active *Fan1* knockdown virus, as evidenced by GFP expression, but low endogenous *Fan1* levels made relative knockdown difficult to ascertain. A combined left striatal and intraperitoneal injection technique was selected for the experimental study.

Fan1 was knocked down by 23% in experimental left striatum and liver throughout the experiment, and there was no effect on right striatum. This level of knockdown is significantly lower than observed *in vitro* an in the toxicity study using the same target sequence.

There was significant CAG repeat expansion in striatum and liver, more so than tail. However, *Fan1* knockdown did not modify expansion rate. There was a suggestion of correlation between *Fan1* expression level and slower expansion in 11 week old striatum and liver, though this was not significant.

It is likely that a greater level of knockdown, or indeed *Fan1* knockout, may be required in order to observe an effect on CAG repeat expansion rate over this experimental timeframe.

Chapter 8 MSH3 modifies somatic instability and disease severity in Huntington's disease and myotonic dystrophy type 1

8.1 Background

8.1.1 Trinucleotide repeat instability in Huntington's and myotonic dystrophy

Huntington's disease (HD) and myotonic dystrophy type 1 (DM1) are autosomal dominant disorders caused by CAG·CTG trinucleotide repeat expansions. HD is characterised by a progressive movement disorder, cognitive impairment and psychiatric symptoms (Bates et al., 2014), and DM1 by myotonia, muscular dystrophy, cognitive impairment, cardiac conduction defects and endocrine dysfunction (Harper, 2001). No disease-modifying treatments are available for either (Meola and Cardani, 2015, Bates et al., 2015b).

HD is caused by a (CAG)n repeat expansion in *HTT* exon 1 and DM1 by a (CTG)n expansion in the 3' untranslated region (UTR) of *DMPK* (Bates et al., 2014, Brook et al., 1992). In both, inherited repeat length is the major determinant of disease course, correlating inversely with the age at onset (AAO) and positively with disease severity. The repeat is unstable and expansion during germline transmission results in genetic anticipation (Bates et al., 2014, Hunter et al., 1992). Consistent with this, there is significant CAG length mosaicism HD patient sperm, which correlates with expansion on transmission (Telenius et al., 1995). In transgenic mice, expansion occurs after meiosis, suggesting DNA replication is not involved (Kovtun and McMurray, 2001). Repeat tracts are also unstable in somatic cells, tending to expand over time, particularly in HD striatum (Kennedy et al., 2003) and DM1 muscle (Ashizawa et al., 1993), the most prominently affected tissues in each disease. Such expansion-biased, age-dependent and tissue-specific somatic instability is thought to contribute to disease onset and progression (Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009, Morales et al., 2012). As this expansion occurs in postmitotic neurons, continues in transgenic mouse cells when the cell cycle is arrested (Gomes-Pereira et al., 2014b), and does not occur in mice when the HD transgene is not expressed (Mangiarini et al., 1997), expansion likely occurs during DNA repair, rather than during replication or transcription.

8.1.2 Modifiers of repeat instability

In CAG·CTG expansion mouse models, the DNA mismatch repair complex MutSβ is essential for repeat expansion, and inactivating or reducing expression of the MutSβ components Msh2 and Msh3 limits expansion events and improves disease phenotype (Dragileva et al., 2009, Tome et al., 2013a, van den Broek et al., 2002, Foiry et al., 2006, Pinto et al., 2013a). Mismatch repair components have also been implicated as genetic modifiers in patients with HD and DM1. A candidate gene association study in DM1 reported a single nucleotide polymorphism (SNP) in *MSH3* that was associated with the rate of somatic expansion (Morales et al., 2016). Genome-wide association studies (GWAS) in HD identified genetic variation in DNA repair genes including *FAN1*, *RRM2B*, *MSH3* and *MLH1*, that modifies disease course, and pathway analyses in each study highlighted sets of DNA repair genes (GeM-HD, 2015, Lee et al., 2017, Hensman Moss et al., 2017b). Such variants have also been shown to influence onset in other CAG expansion polyglutamine diseases (Chapter 3), suggesting a common mechanism operates in conditions caused by repeat expansion (Bettencourt et al., 2016).

The lead variant in a recent GWAS linking *MSH3* with HD progression was the imputed single nucleotide polymorphism (SNP) rs557874766, which nominally results in the Pro67Ala amino acid change at the N-terminus (Hensman Moss et al.,

2017b). However, rs557874766 is located within a 9 bp tandem repeat in exon 1 of *MSH3* and the promoter of the dihydrofolate reductase gene (*DHFR*) on the opposite strand. The 9 bp tandem repeat is polymorphic in copy number (Nakajima et al., 1995, Morales et al., 2016) and sequence (Morales, 2006). The region was first identified by Nakajima et al. (1995), who observed a 3-repeat variant in HeLa cells. In the Japanese population they estimated allele frequencies of 0.603 for 6 repeats, 0.190 for 7 repeats, 0.155 for 3 repeats, 0.043 for 4 repeats and 0.009 for 5 repeats. Additionally, the 500 bp region flanking the *MSH3* repeat is highly polymorphic, containing six SNPs and a 1 bp indel.

8.1.3 MSH3 function

MSH3 complexes with MSH2 to form MutSβ, which binds DNA mismatches incorporated during replication and initiates repair by recruiting the MutLα complex (MLH1/PMS). However, MSH3 has also been shown to bind and stabilise CAG hairpin loops (Owen et al., 2005) and actively repair lesions in non-replicating cells (Rodriguez et al., 2012, Tome et al., 2013a). Its published crystal structure (Acharya et al., 1996) excludes the N-terminal repeat region, in which rs557874766, the variant implicated in HD, is located (Hensman Moss et al., 2017b). *MSH3* is expressed ubiquitously, but it is upregulated in the brains of HD mouse models (Gonitel et al., 2008). Frameshift and compound heterozygous mutations can cause endometrial cancer (Risinger et al., 1996) and recessively inherited colorectal cancer (Adam et al., 2016) respectively, but *Msh3* inactivation is not associated with cancer predisposition (de Wind et al., 1999).

8.2 Aims

Given the complex nature of the 9 bp tandem repeat in exon 1 of *MSH3*, rs557874766 may be an alignment artefact resulting from differences in number and sequence of *MSH3* repeat alleles. The Hensman Moss et al. (2017b) GWAS identified a genomic region associated with disease progression, but the variant with the smallest p-value (the 'lead' SNP) is not necessarily causal (Bush and Moore, 2012, Spain and Barrett, 2015). In order to finely map this association region, this chapter describes targeted Illumina sequencing of the *MSH3* exon 1 region in 218 HD and 247 DM1 subjects. By directly genotyping all the variants within the region, including repeat alleles and flanking variants, with high confidence in multiple patient cohorts, accurate haplotype information was generated (Spain and Barrett, 2015). Though long read sequencing gives significant advantages for complex and repetitive regions, Ilumina MiSeq amplicon sequencing, with a read length of 2 x 300 bp and primers targeting the *MSH3* association region, was sufficient to accurately sequence the 9 bp tandem repeat, and its flanking variants (Pollard et al., 2018). Using lymphoblastoid cells and whole blood RNA-Seq in HD, this chapter studies whether sequence variation at the *MSH3/DHFR* locus influences their expression. A transcriptome-wide association study (TWAS) investigates whether genetic variation in *MSH3* and *FAN1* that is associated with HD disease course also influences expression in brain.

8.3 Methods

8.3.1 Cohorts

The HD cohort was from TRACK-HD, a prospective, observational study with detailed phenotypic data from 218 adults with early HD or premanifest carriers of disease-associated alleles (Tabrizi et al., 2009a).

The DM1 cohort was from OPTIMISTIC, a multicentre randomised clinical trial (van Engelen and Consortium, 2015) that recruited 255 ambulant adult patients affected by severe fatigue and with a genetic diagnosis of DM1.

8.3.2 Progenitor allele length

Molecular diagnosis and genotyping of the HD CAG repeat has traditionally used estimation of PCR fragment size. However, this approach is complicated by the presence of an adjacent polymorphic CCG repeat and provides no information on the presence of variant repeats, flanking sequence variants or on the degree of somatic mosaicism. To overcome these limitations, Ciosi et al. (2018) developed an amplicon-sequencing protocol on the MiSeq (Illumina) platform. Progenitor pure CAG length for HD in this chapter was determined in Ciosi *et al.*, (under review) by MiSeq sequencing (Ciosi et al., 2018). The *HTT* sequence encoding the polyglutamine and polyproline tracts was amplified using MiSeq-compatible PCR primers. The *HTT* locus-specific primers used were HS319F (5'-GCGACCCTGGAAAAGCTGATGA-3') and 33935.5 (5'-AGCAGCGGCTGTGCCTGC-3') which respectively bind 26 bp 5' upstream of the CAG repeat and 26 bp 3' downstream of the CCG repeat. PCR was conducted with the Nextera XT Kit v2 set, and a fraction of each product was pooled and cleaned using AMPure XP beads (Beckman Coulter). Library concentration was measured using a Qubit fluorometer and the dsDNA High Sensitivity (HS) Assay Kit (Qubit), then library size and purity were checked on a Bioanalyzer before accurate quantification by qPCR using the KAPA Library Quantification Kit (KAPABIOSYSTEMS). The library was sequenced on the MiSeq platform.

Five HD subjects were excluded because they were part of a twin pair (n=1) or the progenitor CAG length could not be unambiguously identified (n=4) (Ciosi *et al.,* under review). For the remaining 213 HD subjects, mean CAG repeat length was 43.10 (sd = 2.24), mean AAO was 44.83 (sd = 8.82) and mean progression score was 0.0 (sd = 1.0). CAG length negatively correlated with AAO, with each repeat advancing onset by 3.16 $\pm$ 0.22 years (r^2 = 0.621, p = 2.2E-16).

DM1 progenitor allele length was determined by small pool PCR (van Engelen and Consortium, 2015) (Cumming *et al.*, under review). DM1 patients were tested for CCG repeat interruptions, known *cis*-modifiers of CTG repeat stability and disease phenotype (Cumming et al., 2018) (Cumming et al., under review).

8.3.3 Phenotypes

Two phenotypes were common to both cohorts: AAO and rate of somatic expansion of the pathogenic CAG·CTG repeat. HD AAO represents onset of motor symptoms (Tabrizi et al., 2009a). DM1 AAO was subject self-assessment of the first occurrence of symptoms likely related to DM1 (Cumming *et al.*, under review). Somatic CAG·CTG expansion in blood was previously quantified in both cohorts (Ciosi *et al.* under review; Cumming *et al.*, under review). With sufficient sequencing depth, *HTT* MiSeq data can also be used to quantify the degree of somatic mosaicism of the HTT CAG. For HD MiSeq data, the measure of somatic expansion was the proportion of reads in the sample that correspond to somatic expansions (reads with more CAG repeats than the progenitor allele) relative to the proportion of reads obtained for the progenitor

allele (Ciosi *et al.* under review). This measure correlated with disease course in the Track-HD and Enroll-HD cohorts, suggesting subjects with a greater level of somatic expansion had earlier onset and progressed more rapidly.

$$ln(SM) = \beta_0 + \beta_1 Q^1 + \beta_2 Age + \beta_3 Q^1 * Age$$

Equation 8.1. Regression model of the relationship between disease severity, allele structures and somatic mosaicism in HD, from Ciosi et al. (2018).

β0 (intercept) = -9.5, β1 = 0.18, Q^1 = pure CAG repeat length, β2 = -0.02, β3 = 8.8E-04.

For DM1, the measure of somatic expansion was the difference in number of repeats between the modal allele and the estimated progenitor allele length (Cumming et al., 2018). In both cohorts, relative rate of somatic expansion corresponds to the variation in the measures of somatic expansion that is not explained by age and CAG·CTG repeat length. Positive values reflect a faster rate of somatic expansion.

Two phenotypes were only available for HD; progression score (Hensman Moss et al., 2017b) and gene expression. Progression score was derived for 213 TRACK-HD subjects in Ciosi *et al*., (under review), as described in Hensman Moss et al. (2017b). It measures typical HD progression that is not explained by age and pure CAG repeat length, with positive scores reflecting faster progression. One unit increase in the unified HD progression measure corresponds to an increase of 0·71 units per year (95% CI 0·34-1·08) in the rate of change of UHDRS total motor score (TMS), and an increase of approximately 0·2 units per year (0·12-0·30) in the rate of change of total functional capacity (TFC). Whole blood RNA-Seq was available for a representative sample of 54 premanifest gene carriers and 63 manifest HD subjects from the Track-HD cohort, selected to ensure a range of disease risk and severity. RNA extraction and sequencing were performed as detailed in chapter 4 (Hensman Moss et al., 2017a). Four samples failed quality control for duplication rate over 75%, GC bias or 5' bias, and were removed, leaving 48 premanifest and 61 manifest subjects.

MSH3 expression was measured by Western blot, normalised to actin, and by Taqman qPCR using the Hs00989003_m1 probe set (ThermoFisher), and housekeeping genes *ACTB* (Hs01060665_g1), *ATP5B* (Hs00969569_m1) and *EIF4A2* (Hs00756996_g1) (see Materials and Methods).

8.3.4 Illumina sequencing

MiSeq amplicon sequencing, adapted from Ciosi et al. (2018), was used to genotype the *MSH3* exon 1 repeat and flanking variants (Figure 1). The *MSH3* repeat region was amplified from 10 ng blood genomic DNA using gene-specific primers (forward 5'- AGTTTGGCGCGAAATTGTGG-3', reverse 5'-CTTCCTCCTCCAGCCCTATC-3') with attached Nextera XT Index Kit v2 barcode adaptors (see Appendix). This includes 40 indexes (16 i5 and 24 i7) and can be used to process up to 384 samples per run. The 524 bp amplicon was designed to balance limits of efficient PCR amplification with inclusion of variants in linkage disequilibrium (LD) with rs557874766. A 10 µL PCR reaction included 1 µL DNA, 10% DMSO, 1 µM of each primer, 1 µL of 10X Custom PCR master mix (Thermo Scientific; 45 mM Tris-HCl (pH 8.8), 11 mM $(NH_4)2SO_4$, 4.5 mM $MgCl_2$, 0.113 mg/ml BSA, 4.4 µM EDTA, 1 mM each of dATP, dCTP, dGTP and dTTP), 0.048 % (v/v) 2-mercaptoethanol and 1 U of Taq polymerase (Sigma). Thermal cycling conditions were an initial denaturation of 96°C for 5 min, followed by 30 cycles of 96°C for 45 s, 60°C for 45 s and 70°C for 2 min with a final extension of 65°C for 1 min and 70°C for 10 min. After amplification, 5 µL of each PCR reaction was pooled, purified and concentrated with two cycles of AMPure XP bead clean-up (Beckman Coulter) on a magnetic stand, as described in Ciosi et al. (2018). DNA concentration of the purified library was measured using a Qubit fluorimeter and the dsDNA High Sensitivity (HS) assay kit (Life Technologies). This was used

to prepare a 0.5 ng/µL dilution that could be run on a Bioanalyzer (Agilent) to check library size, ensure primer dimer exclusion and check concentration, aiming for a library of at least 30 µL at 10 nM. The quantity of sequencable molecules was measured by qPCR using the KAPA Library Quantification Kit (KAPABIOSYSTEMS). The concentration of the sequencing library was then calculated and corrected for size using the average fragment size estimated by the Bioanalyzer. The library was diluted to 4 nM, based on the qPCR molarity estimate, and sequenced following Illumina guidelines for an amplicon MiSeq run, using MiSeq Reagent Kit v3 (Illumina) with a cluster density of 1000k cluster/mm^2 but with 5% PhiX spike-in (the PhiX library allows increasing nucleotide diversity during the run and also serves as a sequencing control). The 600 sequencing cycles were run 400 forward, 200 reverse to maximise coverage of the *MSH3* repeat region. Quality control of the sequencing run confirmed >80% of bases had a quality >30. The MiSeq Reporter software (version 2.5.1) was used to demultiplex reads corresponding to the index primer pair used. This outputs the sequencing reads in .fastq format, two (Read 1 and Read 2) for each of the 96 or 384 indexes, as well as two (Read 1 and Read 2) undetermined reads (reads corresponding to the PhiX control library, which is not indexed, and reads for which the indexes could not be identified).

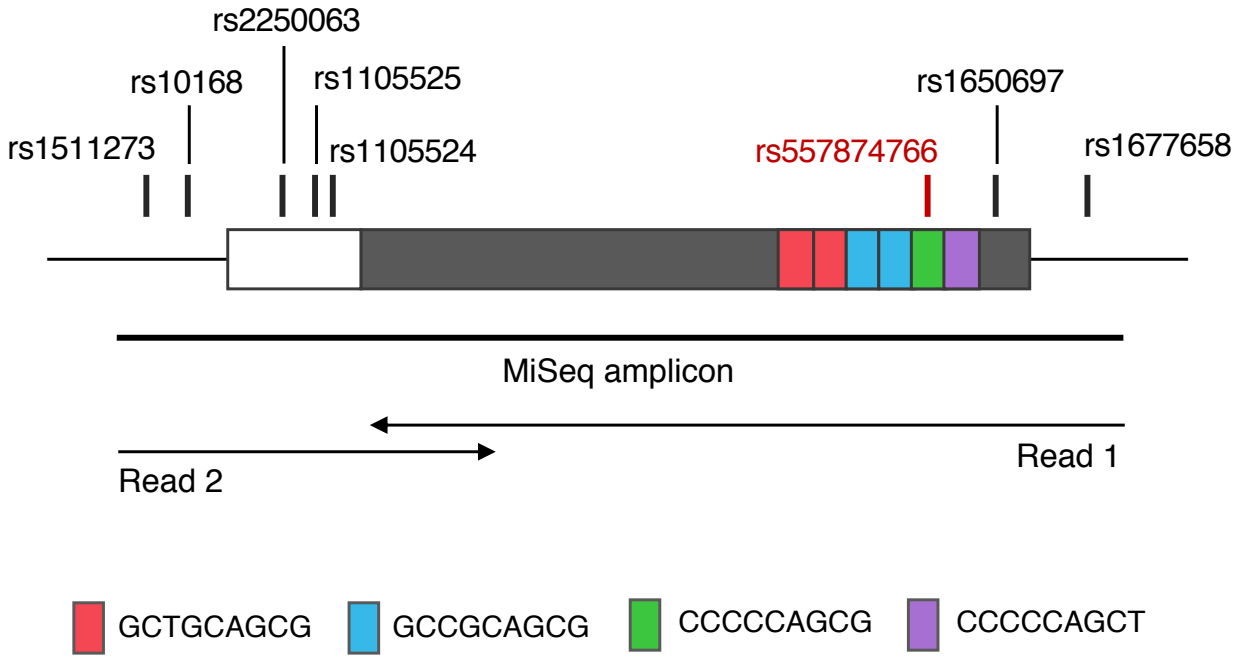


Figure 8.1. Schematic of sequencing design for the MSH3 exon 1 region.
Gene-specific primers with attached MiSeq Illumina barcodes were designed to amplify 534 bp covering MSH3 9 bp tandem repeat region and seven flanking variants. SNP IDs are indicated for each variant, repeat units are coded by coloured boxes, the imputed SNP from GWAS on HD progression is shown in red. Arrows show direction of forward (Read 1) and reverse (Read 2) reads during MiSeq Illumina sequencing, 400 nucleotides were sequenced in a forward read and 200 nucleotides in a reverse read.

MiSeq runs were quality controlled by running the 'PhiX.R1andR2.QC.400x200bp.run' workflow on undetermined.fastq files, which correspond to PhiX reads and lack barcodes, with the PhiX reference sequence (see Appendix). This confirmed >80% of bases had a Phred quality >30.

8.3.5 Bioinformatic analyses

Genotyping was conducted on the University of Glasgow Galaxy platform (heighliner.cvr.gla.ac.uk). Paired-end reads were merged with Pear and aligned with BWA-MEM to 84 references corresponding to potential 9 bp repeat alleles (see

Appendix), followed by variant calling using the Naive Variant Caller (NVC). For repeat homozygotes, haplotypes were confirmed from .sam files using Tablet (Milne et al., 2013). The Galaxy workflow is available in the Appendix and at https://www.myexperiment.org/workflows/5087.html. Conservation analysis used PhastCons and PhyloP (UCSC), with species sequence alignment in Clustal Omega.

8.3.6 Transcriptome-wide association study (TWAS)

The method of Gusev et al. (2016) was used to test for association between phenotype and gene expression in control dorsolateral prefrontal cortex from the CommonMind Consortium (n=452) (CMC, 2017) using GWAS summary statistics from the Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium GWAS of AAO (n = 4082) (GeM-HD, 2015) and the combined TRACK-HD and REGISTRY GWAS of HD progression (n = 2078) (Hensman Moss et al., 2017b).

8.3.7 Statistical analyses

Linear regression modelling of genotype-phenotype correlation was conducted in R (R Core Team, 2013). An additive genetic model was used to score genotypes. For AAO, analyses controlled for CAG·CTG repeat length in HD and DM1, and for repeat interruptions in DM1 (Table 1). Meta-analysis of somatic expansion and AAO in HD and DM1 was conducted with METAL (Willer et al., 2010). PLINK 1.07 (Purcell et al., 2007) was used to derive allele frequencies, Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD). Haplotype relationships were visualised as a network using median joining on NETWORK (Bandelt et al., 1999).

	Model	Adjusted r^2	Model *p*	Parameter		Coefficient	Standard error	*t*-statistic	Parameter *p*
1a	$AAO = \beta 0 + \beta 1\ log_{10}(CAG.CTG) + \beta 2\ Variant\ Repeats$	0.24	3.35×10^{-14}	*Intercept*	$\beta 0$	86.63	7.73	11.21	$< 2 \times 10^{-16}$
	OPTIMISTIC, n = 222			$log_{10}(CAG.CTG)$	$\beta 1$	-26.5	3.3	-8.04	5.50×10^{-14}
				Variant Repeats	$\beta 2$	13.85	3.2	4.33	2.23×10^{-05}
1b	$log_{10}(SE) = \beta 0 + \beta 1\ log_{10}(CAG.CTG) + \beta 2\ Variant\ Repeats$	0.581	$< 2.2 \times 10^{-16}$	*Intercept*	$\beta 0$	-0.96	0.18	-5.47	1.14×10^{-07}
	OPTIMISTIC, n = 247			$log_{10}(CAG.CTG)$	$\beta 1$	1.38	0.07	18.46	$< 2 \times 10^{-16}$
				Variant Repeats	$\beta 2$	-0.38	0.07	-5.31	2.47×10^{-07}
2a	$ln(AAO) = \beta 0 + \beta 1\ CAG.CTG$	0.661	$< 2.2 \times 10^{-16}$	*Intercept*	$\beta 0$	6.59	0.18	37.03	$< 2 \times 10^{-16}$
	TRACK-HD, n = 130			*CAG.CTG*	$\beta 1$	-0.06	0	-15.88	$< 2 \times 10^{-16}$

Table 8.1. Regression models of the relationships between allele structures, relative rate of somatic expansion and disease phenotypes in Huntington's disease and myotonic dystrophy type 1.

The table shows the adjusted squared coefficient of correlation (adjusted r^2), and statistical significance (p) for each model, and the coefficient, standard error, t statistic and statistical significance (p), associated with each parameter in the model. The coefficient provides an indication of the relative weight of the contribution of each parameter to the model and its associated standard error. The t statistic and corresponding p value provide an indication of the statistical significance that the parameter is adding to the explanatory power to the model. AAO: age at onset of HD and DM1. CAG·CTG: inherited repeat allele length in HD and DM1. Variant repeats – presence/absence of repeat interruptions in DM1 expanded repeat allele. SE: relative rate of somatic expansions. Note that model for relative rate of somatic expansion, not shown here, was obtained from (Ciosi et al. under review).

8.4 Contributions

This study was conceived by Professors Sarah Tabrizi (UCL) and Darren Monckton (University of Glasgow). The Illumina sequencing protocol was developed by Marc Ciosi and Vilija Lomeikaite (University of Glasgow). Sequencing and analysis in the HD cohort was conducted by Michael Flower and in the myotonic dystrophy type 1 cohort by Vilija Lomeikaite. Somatic instability in the DM1 cohort was determined by Fernando Morales (University of Glasgow), and for the HD cohort DNA was prepared by Michael Flower and instability was measured by Marc Ciosi. HD patient blood RNA was prepared by Davina Hensman Moss, sequenced at Biorep, aligned by Kitty Lo (University of Sydney) and Vincent Plagnol (UCL), and analysed by Michael Flower. Statistical analyses were conducted by Michael Flower. The transcriptome-wide association study (TWAS) was conducted by Prof Peter Holmans (Cardiff University). These results were included in a manuscript written by Michael Flower, which has been accepted for publication in *Brain*.

8.5 Results

8.5.1 Rs557874766 is an alignment artefact

In order to better understand the association between the variation in *MSH3* and trinucleotide repeat disease phenotype, the exon 1 region was sequenced in cohorts of HD and DM1 patients. Sanger sequencing shows the 9 bp tandem repeat is polymorphic and varies in the repeat number and sequence, with heterozygosity of 0.57 (Nakajima et al., 1995).

Here, sixteen *MSH3* repeat alleles were observed, differing in sequence and length from three to nine repeats (Figure 2A and Table 2). Alleles contained combinations of five types of repeat units, with coding potential for proline or alanine (Figure 2A). They were numbered by repeat length, and suffixed alphabetically by frequency i.e. '3a' represents the most common three-repeat allele.

The commonest allele in both cohorts, 6a (Figure 2B), corresponds to the human reference sequence (NC_000005.10, GRCh38.p12). Illumina sequencing revealed that rs557874766 (Hensman Moss et al., 2017b) was not a SNP, but an alignment artefact resulting from the complex 9 bp repeat sequence (Figure 2C). Individuals with the rs557874766 minor allele instead carry a three-repeat allele, 3a, the second most common allele observed in both cohorts. Two HD subjects imputed as homozygous for the rs557874766 major allele were determined to be heterozygous for the 3a repeat allele by both Illumina and Sanger sequencing (Figure 3), highlighting the importance of directly genotyping such complex loci. In conclusion, the rs557874766 does not exist in the form of a SNP and results from incorrect alignment of the 3a allele to the reference 6a allele (Figure 2C).

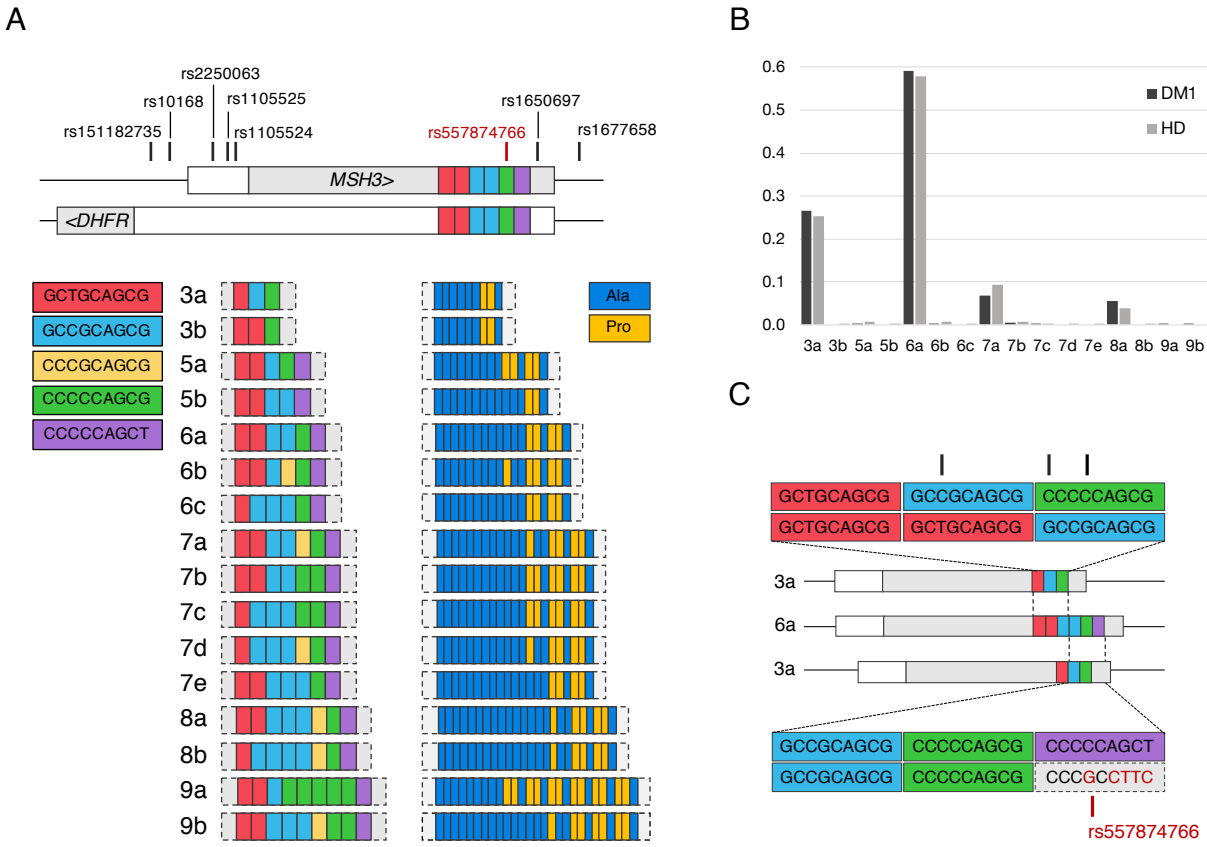


Figure 8.2. MSH3/DHFR 9bp tandem repeat allele structure and frequency observed in HD and DM1 cohorts.
(A) *Schematic representation of the 9 bp tandem repeat alleles observed in this study and their coding potential. Repeat units are colour-coded by DNA and amino acid sequence. Location of the repeat and flanking variants in relation to MSH3/DHFR locus are shown in the top panel. This locus contains overlapping MSH3 exon 1 and DHFR promoter regions. For both MSH3 and DHFR, the 5'-untranslated region is shown in white and coding sequence in light grey. The direction of transcription is indicated by arrows for each gene.* ***(B)*** *Repeat allele frequencies observed in HD and DM1. Four common alleles, 3a, 6a, 7a and 8a, are observed in HD and DM1 cohorts at similar frequencies.* ***(C)*** *Schematic showing potential misalignments of 3a and 6a alleles, resulting in the apparent SNP rs557874766, shown in red on the lower alignment. Black marks in the top alignment represent mismatches that could be created in a similar manner as rs557874766, by misalignment of the 3a and 6a repeat alleles.*

MSH3 **repeat allele**	**Number of repeats**	**Repeat Sequence**	**Allele Frequency**	
			HD	**DM1**
3a	3	(GCTGCAGCG)1(GCCGCAGCG)1(CCCGCAGCG)0(CCCCCAGCG)1(CCCCCAGCT)0	0.251	0.266
3b	3	(GCTGCAGCG)2(GCCGCAGCG)0(CCCGCAGCG)0(CCCCCAGCG)1(CCCCCAGCT)0	0.002	0.000
5a	5	(GCTGCAGCG)2(GCCGCAGCG)1(CCCGCAGCG)0(CCCCCAGCG)1(CCCCCAGCT)1	0.007	0.004
5b	5	(GCTGCAGCG)2(GCCGCAGCG)2(CCCGCAGCG)0(CCCCCAGCG)0(CCCCCAGCT)1	0.002	0.000
6a	6	(GCTGCAGCG)2(GCCGCAGCG)2(CCCGCAGCG)0(CCCCCAGCG)1(CCCCCAGCT)1	0.585	0.592
6b	6	(GCTGCAGCG)2(GCCGCAGCG)1(CCCGCAGCG)1(CCCCCAGCG)1(CCCCCAGCT)1	0.007	0.002
6c	6	(GCTGCAGCG)1(GCCGCAGCG)3(CCCGCAGCG)0(CCCCCAGCG)1(CCCCCAGCT)1	0.002	0.000
7a	7	(GCTGCAGCG)2(GCCGCAGCG)2(CCCGCAGCG)1(CCCCCAGCG)1(CCCCCAGCT)1	0.092	0.069
7b	7	(GCTGCAGCG)2(GCCGCAGCG)2(CCCGCAGCG)0(CCCCCAGCG)2(CCCCCAGCT)1	0.005	0.006
7c	7	(GCTGCAGCG)1(GCCGCAGCG)3(CCCGCAGCG)0(CCCCCAGCG)2(CCCCCAGCT)1	0.002	0.002
7d	7	(GCTGCAGCG)1(GCCGCAGCG)3(CCCGCAGCG)1(CCCCCAGCG)1(CCCCCAGCT)1	0.002	0.000
7e	7	(GCTGCAGCG)2(GCCGCAGCG)3(CCCGCAGCG)0(CCCCCAGCG)1(CCCCCAGCT)1	0.002	0.000
8a	8	(GCTGCAGCG)2(GCCGCAGCG)3(CCCGCAGCG)1(CCCCCAGCG)1(CCCCCAGCT)1	0.036	0.055
8b	8	(GCTGCAGCG)1(GCCGCAGCG)4(CCCGCAGCG)1(CCCCCAGCG)1(CCCCCAGCT)1	0.002	0.000
9a	9	(GCTGCAGCG)2(GCCGCAGCG)1(CCCGCAGCG)0(CCCCCAGCG)5(CCCCCAGCT)1	0.000	0.002
9b	9	(GCTGCAGCG)2(GCCGCAGCG)3(CCCGCAGCG)1(CCCCCAGCG)2(CCCCCAGCT)1	0.000	0.002

Table 8.2. MSH3 9 bp tandem repeat alleles observed in HD and DM1 cohorts.
Allele frequency in each cohort is given.

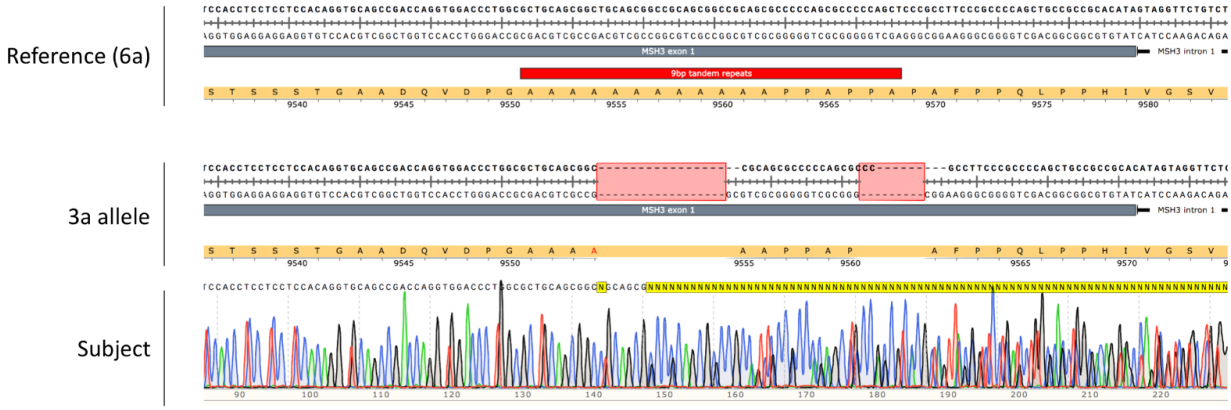


Figure 8.3. Representative Sanger sequencing of a 3a heterozygote.
Two HD samples genotyped as homozygous for the reference allele rs557874766 were subsequently found to have one 3a allele each by MiSeq Illumina sequencing. Sanger sequencing confirmed heterozygosity at the 9 bp tandem repeat, consistent with the MiSeq sequencing result (representative trace shown). On top is the human reference sequence (GRCh38), then the 3a allele sequence (note this is aligns to the reference as separate 18 and 9 bp deletions, light red). The 9 bp tandem repeat is marked in relation to the reference sequence in red.

The *MSH3* exon 1 repeat region is poorly conserved between species, with mean scores of 0.29 (SD 0.41) and 0.25 (SD 0.91) in PhastCons and PhyloP respectively (see Appendix). Sequence alignment of 20 mammalian reference genomes showed most have two repeats (Figure 8.4). Together with a four- and a five-repeat allele, the 3a allele has been observed in gorillas and chimpanzees, suggesting 3a is an ancestral allele in humans (Morales, 2006).

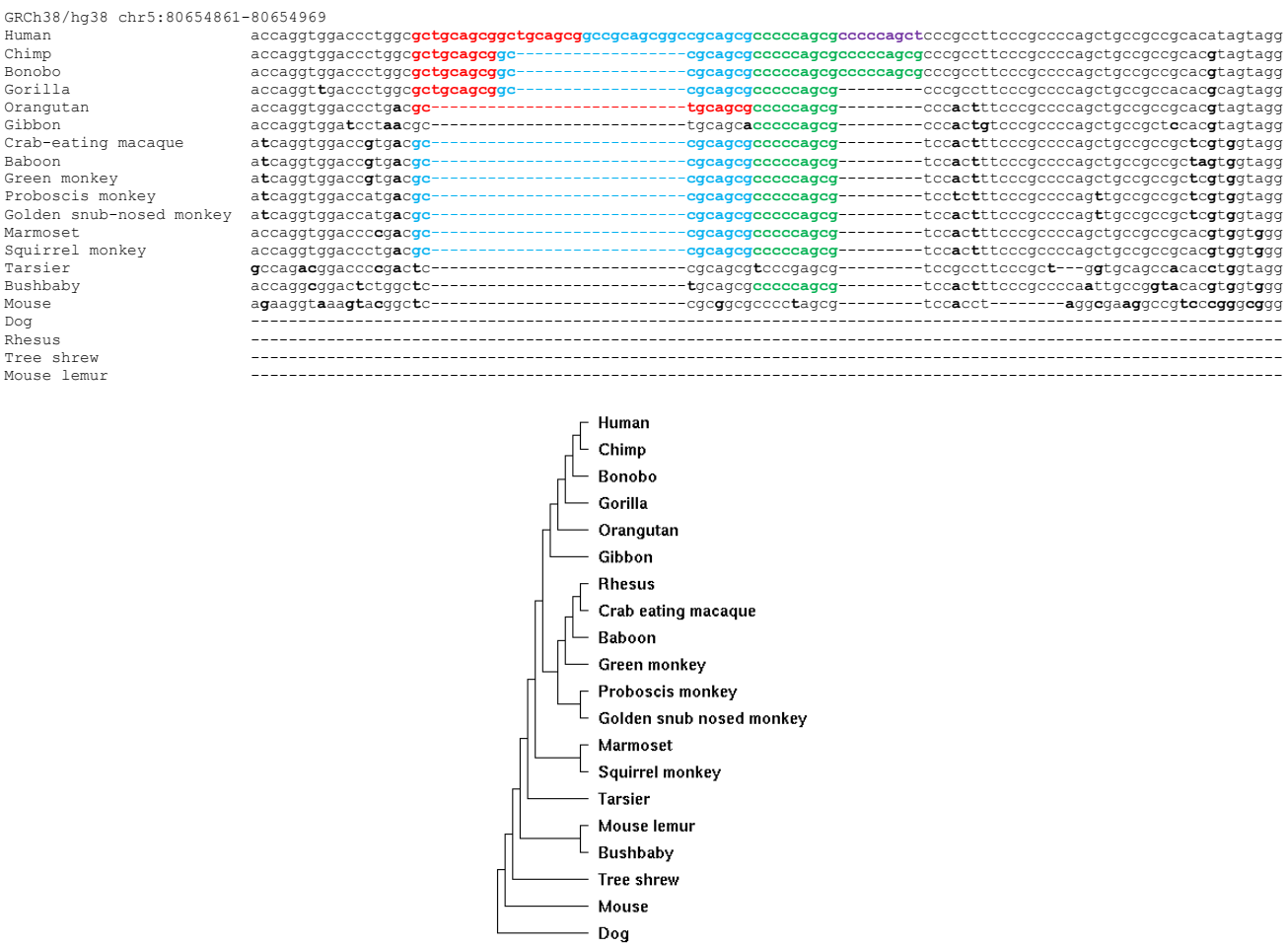


Figure 8.4. The MSH3 N-terminal region is poorly conserved between species.

Alignment of MSH3 orthologues of 20 mammals (17 primates). Sequences from Ensembl, aligned in Clustal. Most mammals have two repeats (blue and green). Gorillas have the three-repeat allele observed in slow progressing HD subjects, and chimpanzees have four repeats. Rodents and canines lack the 9 bp tandem repeat region

8.5.2 Variants at the *MSH3/DHFR* locus are associated with the rate of somatic expansion and disease phenotypes in HD and DM1

The 3a allele correlated negatively with relative rate of somatic expansion in HD subjects (p=0.032) and showed similar effect direction, though above nominal significance, in DM1 (p=0.053) (Fig. 8.5, Table 8.3). Additionally, 3a was associated with delayed AAO by 1.05 years (p=0.0029) and slower progression in HD by 0.52 units ($p=3.86\times10^{-7}$), which corresponds to 0.37 and 0.10 units per year on the UHDRS total motor score and total functional capacity respectively. In DM1, the association between 3a and AAO showed a consistent effect direction, approaching significance (p=0.061). In meta-analysis, 3a was significantly associated with relative rate of somatic expansion (p=0.004) and AAO (p=0.003) in HD and DM1. Detailed analysis of the relationship between repeat alleles and phenotypes (Table 8.4) showed that the 3a allele accounts for the reduced somatic expansion rate, delayed onset and slower progression observed in HD. In DM1, the number of seven-repeat alleles was associated with reduced expansion rate (Table 8.4).

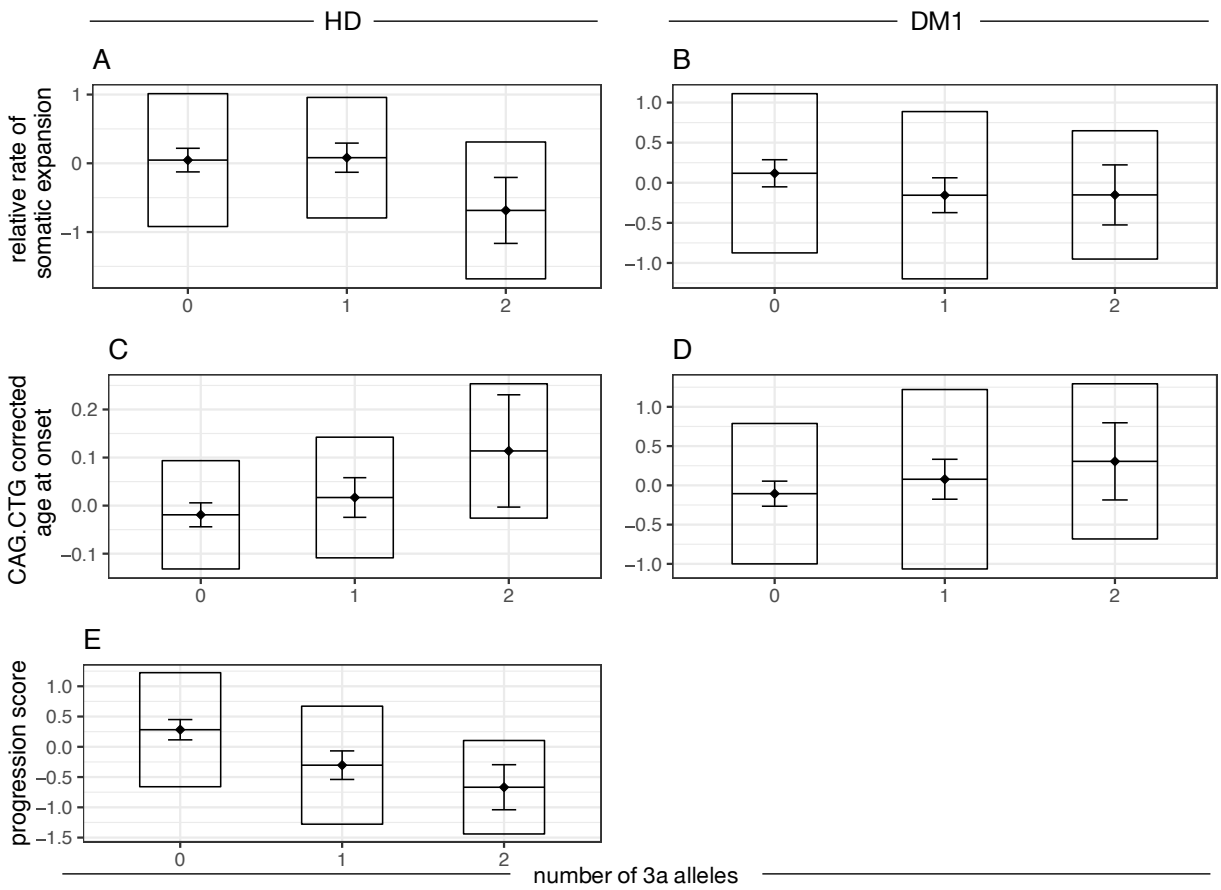


Figure 8.5. The number of MSH3 3a repeat alleles is associated with HD and DM1 phenotypes.

*Boxplots for three measures of disease phenotype are shown: rate of somatic expansion corrected for the inherited CAG·CTG length in HD **(A)** and for the inherited CAG·CTG length and variant repeats in DM1 **(B)**; age at onset corrected for the inherited CAG·CTG length in HD **(C)** and DM1 **(D)**; progression score in HD **(E)**. For each dataset, the diamond and horizontal line spanning the diamond indicate the mean, the box the standard deviation and the whiskers the 95% confidence intervals of the mean.*

MSH3 repeat allele	Regression with relative rate of somatic expansion												Regression with residual variation in age at onset												Regression with progression score			*MSH3* expression (blood)			*DHFR* expression (blood)			*MSH3* expression (cortex, imputed)			*DHFR* expression (cortex, imputed)		
	HD			DM1			Meta-analysis ($DM1_{OPTIMISTIC}$ + HD_{TRACK})		Meta-analysis ($DM1_{OPTIMISTIC}$ + $DM1_{CostaRica}$)		Meta-analysis ($DM1_{OPTIMISTIC}$ + $DM1_{CostaRica}$ + HD_{TRACK})		HD			DM1			Meta-analysis ($DM1_{OPTIMISTIC}$ + HD_{TRACK})		Meta-analysis ($DM1_{OPTIMISTIC}$ + $DM1_{CostaRica}$)		Meta-analysis ($DM1_{OPTIMISTIC}$ + $DM1_{CostaRica}$ + HD_{TRACK})		HD			HD			HD			HD			HD		
	β	*p*	r2	β	*p*	r2	Z-score	*p*	Z-score	*p*	Z-score	*p*	β	*p*	r2	β	*p*	r2	Z-score	*p*	Z-score	*p*	Z-score	*p*	β	*p*	r2	β	*p*	r2	β	*p*	r2	β	*p*	r2	β	*p*	r2
3a	-0.217	**0.032**	0.017	-0.193	0.053	0.011	-2.876	**4.02E-03**	-2.87	**4.16E-03**	-3.58	**3.46E-04**	0.052	**2.96E-03**	0.061	2.277	0.061	0.011	2.997	**2.72E-03**	2.06	**3.99E-02**	3.02	**2.57E-03**	-0.516	**3.86E-07**	0.112	-0.167	0.139	0.012	-0.400	**2.48E-04**	0.116	-1.363	**4.55E-75**	0.801	-0.319	**2.49E-03**	0.039
3b																																							
5a																																							
5b																																							
6a	0.077	0.393	-0.001	0.238	**8.86E-03**	0.024	2.498	**0.012**					-0.028	0.078	0.017	-2.981	**7.04E-03**	0.028	-1.252	0.211					0.329	**3.18E-04**	0.056	-0.126	0.189	0.007	0.362	**8.59E-05**	0.134	0.609	**6.20E-12**	0.200	0.403	**1.25E-05**	0.084
6b																																							
6c																																							
7a	0.112	0.485	-0.002	-0.468	**0.011**	0.022	-1.393	0.164					-0.019	0.519	-0.005	3.983	0.076	0.010	1.779	0.075					-0.046	0.781	-0.004	0.555	**8.55E-04**	0.096	-0.137	0.421	-0.003	0.893	**3.60E-08**	0.132	-0.576	**5.66E-04**	0.051
7b																																							
7c																																							
7d																																							
7e																																							
8a	0.302	0.242	0.002	0.175	0.393	-0.001	1.421	0.155					-2.13E-02	0.615	-0.006	0.426	0.861	-0.004	0.146	0.884					0.705	**7.87E-03**	0.029	0.696	**8.26E-03**	0.096	-0.116	0.663	-0.008	1.119	**2.19E-05**	0.079	0.443	0.102	0.008
8b																																							
9a																																							
9b																																							

Table 8.3. MSH3 9 bp tandem repeat alleles and their association with phenotypes in DM1 and HD.

An additive genetic model was used to score repeat genotypes and run linear regression analysis. Relative rate of somatic expansion and age at onset were corrected for CAG·CTG length (DM1 and HD) and variant repeats (DM1).

Repeat covariate	Regression with relative rate of somatic expansion						Residual variation in age at onset						Progression score		MSH3 expression		DHFR expression	
	HD		DM1		Meta-analysis		HD		DM1		Meta-analysis		HD		HD		HD	
	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*	β	*p*
number of 3-repeat alleles	-0.215	0.031	-0.192	**0.053**	-0.207	**3.71E-03**	0.415	**3.28E-03**	0.197	**0.060**	0.272	**1.15E-03**	-0.504	**5.63E-07**	-0.160	0.146	-0.393	**2.33E-04**
homozygous 3-repeat genotype	-0.729	**1.12E-03**	-0.158	0.497	-0.456	**5.66E-03**	0.919	**7.32E-03**	0.343	0.164	0.536	7.20E-03	-0.702	2.49E-03	-0.530	0.035	-0.766	1.93E-03
number of 7-repeat alleles	0.146	0.333	-0.559	**1.42E-03**	-0.160	0.170	-0.171	0.436	0.327	**0.076**	0.121	0.393	0.009	0.954	0.535	1.09E-03	-0.088	0.599
number of 8-repeat alleles	0.223	0.344	0.174	0.394	0.197	0.207	-0.175	0.615	0.038	0.859	-0.019	0.914	0.596	0.014	0.698	8.26E-03	-0.116	0.663
number of 7 or 8-repeat alleles	0.138	0.166	-0.266	0.058	-0.035	0.720	-0.180	0.343	0.218	0.131	0.072	0.532	0.196	0.153	0.664	**4.53E-06**	-0.110	0.469
sum of repeat lengths across both alleles	0.069	0.019	0.045	0.119	0.058	**5.93E-03**	-0.118	**4.48E-03**	-0.039	0.200	-0.066	6.97E-03	0.152	**3.34E-07**	0.089	0.007	0.102	1.76E-03

Table 8.4. Detailed investigation of the role of repeat alleles in phenotypic modification.

Note that age at onset and somatic expansion residuals are standardised, to facilitate meta-analysis. Beta and p-value are given for each covariate separately. Covariates in bold are present in the best-fitting model (forward stepwise regression). Where two or more covariates are highlighted, it is impossible to tell which is driving the association, except for DM1 where both the number of 7-repeat and the number of 3-repeat alleles reduce expansion and increase age at onset. Association between DHFR and number of 3-repeat alleles remains significant after correcting for MSH3 expression (p=0.000751). Association between MSH3 and number of 7 or 8-repeat alleles remains significant after correcting for DHFR expression (p=1.30E-07). The 6-repeat allele was the baseline to which the other alleles were compared.

In addition to testing repeat allele effects, we also assessed correlation between flanking SNP genotypes and disease phenotypes. All the flanking variants were in HWE (Table 8.5) and in strong LD with each other (Fig. 6B). Three variants (rs151182735, rs10168 and rs2250063) were in nearly complete LD with the 3a allele, and as such were as significantly associated with phenotypes (Fig. 8.6A, Table 8.6). All three are non-coding variants 5' to the repeat and their alternative alleles are associated with reduced *MSH3* and *DHFR* expression in the prefrontal cortex (CMC, 2017) and in multiple tissues in GTEx (GTEx, 2015) (Appendix). Three SNPs, rs1105524, rs1650697 and rs1677658, also correlated with some phenotypes, though not uniformly (Fig. 6A and Table 6). Rs1105524 and rs1677658 are non-coding variants, whereas rs1650697 corresponds to Ile79Val. All three are expression quantitative trait loci (eQTL) for *MSH3* and *DHFR* in the prefrontal cortex (CMC, 2017) and in multiple tissues in GTEx (Appendix).

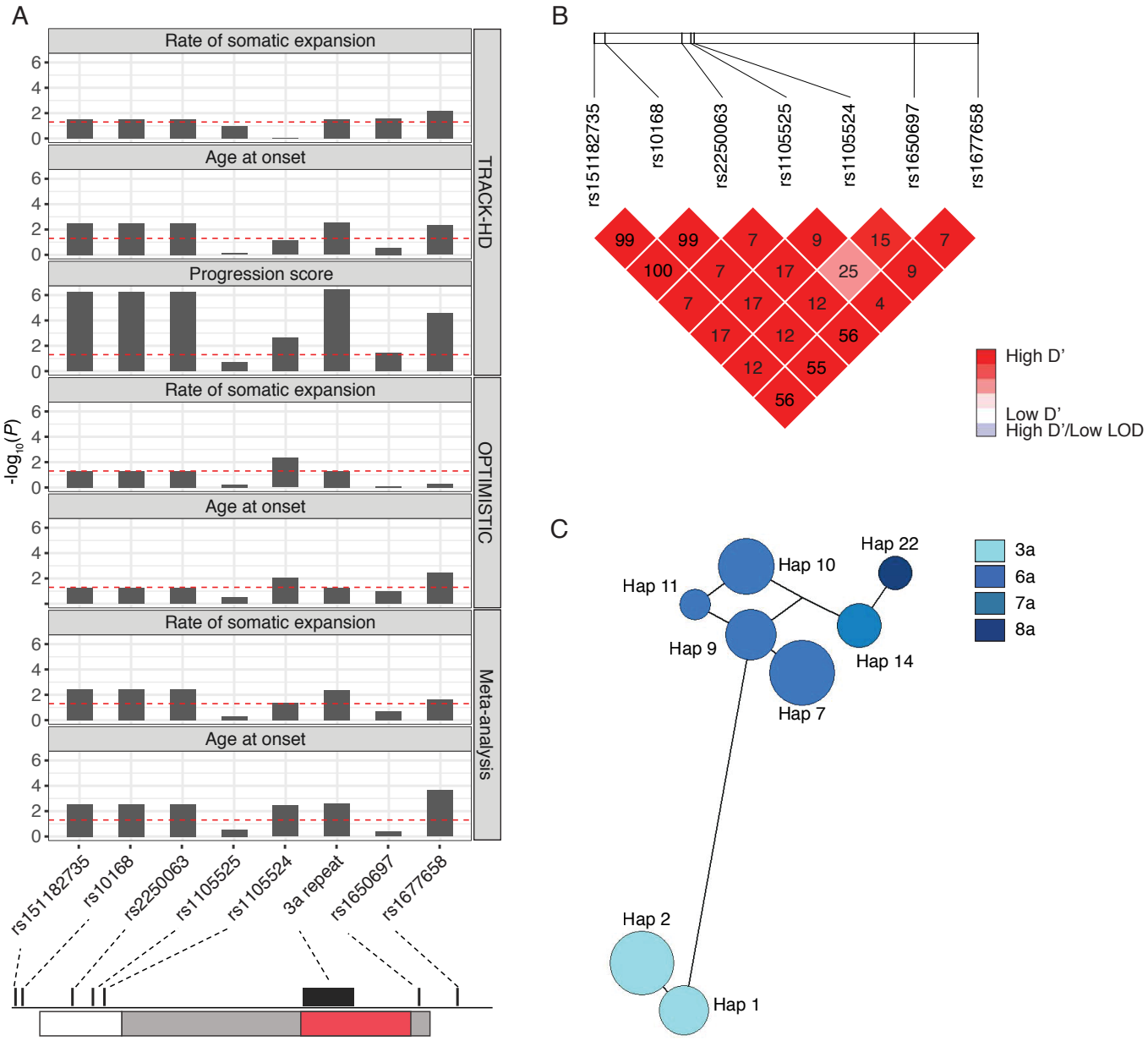


Figure 8.6. Variants at the MSH3/DHFR locus are associated with phenotypes in HD and DM1.

(A) *Bar charts showing associations between variant genotypes and disease phenotypes: relative rate of somatic expansion and age at onset corrected for the CAG·CTG length and progression score for HD (TRACK-HD), rate of somatic expansion and age at onset corrected for the CAG·CTG length and repeat interruptions for DM1 (OPTIMISTIC), and rate of somatic expansion and age at onset in the meta-analysis of HD and DM1. Each bar represents association for a single variant. Red dotted line in plot panels represents the p=0.05 significance threshold. Variant location in relation to the MSH3 exon 1 region is shown in the bottom panel; white box – 5' untranslated region, grey – coding sequence, red – MSH3 repeat region, intron is shown by a black line.* ***(B)*** *Linkage disequilibrium heatmap for the seven variants flanking the MSH3 repeat. Colour intensity represents the D' value for each SNP pair. R2 values are indicated in text for each variant pair.* ***(C)*** *Haplotype network for eight haplotypes with frequency > 0.035 observed at the MSH3 exon 1 region. Circles represent different haplotypes. The size of the circle is proportional to the number of individuals with a particular haplotype. Each haplotype is connected with the most similar haplotype by a line. Length of the line represents the number of genotypes that are different between each two haplotypes. Circles are colour coded according to the repeat allele found on the haplotype.*

SNP ID	Chromosomal position	Reference allele	Alternative allele	Minor allele	Minor allele frequency		HWE (HD+DM1)	LD with 3a repeat allele (HD+DM1)
					HD	DM1		
							p	r2
rs151182735	5:80654571	GC	G	G	0.255	0.266	0.0511	0.995
rs10168	5:80654584	C	T	T	0.255	0.266	0.0511	0.995
rs2250063	5:80654678	C	T	T	0.255	0.266	0.0511	0.995
rs1105525	5:80654689	C	T	T	0.164	0.171	0.2478	0.081
rs1105524	5:80654693	A	G	A	0.343	0.313	0.0054	0.192
rs746491510	5:80654720	G	T	T	0.002	0.002		0.003
rs6151597	5:80654748	G	A	A	0.002	0.000		0.001
rs1650697	5:80654962	A	G	A	0.264	0.268	0.654	0.143
rs943394665	5:80655025	A	T	T	0.002	0.002		0.010
rs1677658	5:80655040	G	T	T	0.174	0.159	0.0533	0.610

Table 8.5. MSH3 exon 1 region variants.

Linkage disequilibrium (LD) with the 3a allele is given in r^2. Note that correct alignment of the repeat region shows rs1105524 and rs1650697 alternative allele frequencies are >0.5. HWE – Hardy-Weinberg equilibrium.

SNP ID	Regression with relative rate of somatic expansion												Regression with residual variation in age at onset												Regression with progression score			*MSH3* expression (blood)			*DHFR* expression (blood)			*MSH3* expression (cortex, imputed)			*DHFR* expression (cortex, imputed)		
	HD			DM1			Meta-analysis ($DM1_{OPTIMISTIC}$ + HD_{TRACK})		Meta-analysis ($DM1_{OPTIMISTIC}$ + $DM1_{CostaRica}$)		Meta-analysis ($DM1_{OPTIMISTIC}$ + $DM1_{CostaRica}$ + HD_{TRACK})		HD			DM1			Meta-analysis ($DM1_{OPTIMISTIC}$ + HD_{TRACK})		Meta-analysis ($DM1_{OPTIMISTIC}$ + $DM1_{CostaRica}$)		Meta-analysis ($DM1_{OPTIMISTIC}$ + $DM1_{CostaRica}$ + HD_{TRACK})		HD			HD			HD			HD			HD		
	β	*p*	r2	β	*p*	r2	Z-score	*p*	Z-score	*p*	Z-score	*p*	β	*p*	r2	β	*p*	r2	Z-score	*p*	Z-score	*p*	Z-score	*p*	β	*p*	r2	β	*p*	r2	β	*p*	r2	β	*p*	r2	β	*p*	r2
rs151182735	-0.215	**0.031**	0.018	-0.193	0.053	0.011	-2.885	**0.004**					0.051	**3.28E-03**	0.060	2.277	0.061	0.011	2.981	**2.87E-03**					-0.504	**5.63E-07**	0.109	-0.160	0.146	0.011	-0.393	**2.33E-04**	0.118	-1.356	**2.22E-77**	0.811	-0.320	**2.15E-03**	0.040
rs10168	-0.215	**0.031**	0.018	-0.193	0.053	0.011	-2.885	**0.004**	-2.866	**0.004**	-3.585	**3.37E-04**	0.051	**3.28E-03**	0.060	2.277	0.061	0.011	2.981	**2.87E-03**	2.055	**0.03988**	3.00	**2.68E-03**	-0.504	**5.63E-07**	0.109	-0.160	0.146	0.011	-0.393	**2.33E-04**	0.118	-1.356	**2.22E-77**	0.811	-0.320	**2.15E-03**	0.040
rs2250063	-0.215	**0.031**	0.018	-0.193	0.053	0.011	-2.885	**0.004**					0.051	**3.28E-03**	0.060	2.277	0.061	0.011	2.981	**2.87E-03**					-0.504	**5.63E-07**	0.109	-0.160	0.146	0.011	-0.393	**2.33E-04**	0.118	-1.356	**2.22E-77**	0.811	-0.320	**2.15E-03**	0.040
rs1105525	0.212	0.101	0.008	-0.071	0.566	-0.003	0.692	0.489					-0.007	0.739	-0.007	1.456	0.332	-2.53E-04	1.064	0.287					0.175	0.192	0.003	0.124	0.394	-0.003	0.010	0.945	-0.010	0.671	**3.22E-07**	0.114	0.572	**1.86E-05**	0.080
rs1105524	0.006	0.950	-0.005	-0.261	**4.63E-03**	0.028	-2.019	**0.044**					0.028	0.074	0.017	2.928	**8.57E-03**	0.027	2.914	**3.57E-03**					-0.290	**2.15E-03**	0.040	0.160	0.122	0.014	-0.442	**7.78E-06**	0.172	-0.278	**3.58E-03**	0.035	-0.047	0.628	-0.004
rs746491510																																							
rs6151597																																							
rs1650697	-0.232	**0.026**	0.019	0.03183	0.753	-0.004	-1.280	0.201					0.019	0.284	0.001	-2.021	0.100	0.008	-0.803	0.422					-0.225	**0.038**	0.016	-0.494	**7.01E-06**	0.174	0.208	0.069	0.023	-1.078	**2.69E-30**	0.465	-0.083	0.452	-0.002
rs943394665																																							
rs1677658	-0.315	**0.007**	0.030	-0.076	0.525	-0.002	-2.297	**0.022**	-2.164	**0.030**	-3.325	**8.85E-04**	0.058	**0.005**	0.055	4.250	**3.43E-03**	0.271	3.684	**2.30E-04**	2.614	**0.008957**	3.352	**0.0008023**	-0.501	**2.50E-05**	0.077	-0.268	**0.028**	0.038	-0.390	**1.12E-03**	0.091	-1.333	**2.64E-40**	0.571	-0.043	0.728	-0.004

Table 8.6. MSH3 exon 1 region variants and the association of their alternative alleles with phenotypes in DM1 and HD.

Linkage disequilibrium (LD) with the 3a allele is given in R2. An additive genetic model was used to score variant genotypes and run linear regression analysis. Relative rate of somatic expansion and age at onset were corrected for CAG·CTG length (DM1 and HD) and variant repeats (DM1). MSH3 and DHFR expression are derived from RNA-Seq in HD whole blood. Note that correct alignment of the repeat region shows rs1105524 and rs1650697 alternative allele frequencies are >0.5. HWE – Hardy-Weinberg equilibrium.

Phenotype	Dataset	rs151182735		rs10168		rs2250063		rs1105525		rs1105524		rs1650697		rs1677658	
		β	p	β	p	β	p	β	p	β	p	β	p	β	p
Relative rate of somatic expansion	HD+DM1		NA		NA		NA	-0.014	0.883	-0.059	0.42	-0.015	0.849	-0.015	0.914
	HD	0.043	0.773	0.043	0.773	0.043	0.773	0.135	0.319	0.113	0.251	-0.159	0.151	-0.134	0.382
	DM1		NA		NA		NA	-0.185	0.142	-0.128	0.222	-0.04	0.738	0.257	0.147
Residual variation in age at onset	HD+DM1		NA		NA		NA	0.162	0.121	0.165	0.049	-0.2	**0.029**	0.375	**0.015**
	HD		NA		NA		NA	0.071	0.673	0.09	0.505	0.005	0.976	0.23	0.382
	DM1		NA		NA		NA	0.244	0.068	0.163	0.145	-0.256	0.043	0.42	0.028
Progression score	HD		NA		NA		NA	-0.008	0.95	-0.107	0.285	-0.05	0.646	-0.054	0.789
MSH3 expression	HD	-0.091	0.367	-0.091	0.367	-0.091	0.367	0.174	0.186	-0.01	0.92	-0.271	0.076	-0.199	0.075
DHFR expression	HD		NA		NA		NA	-0.206	0.156	-0.345	**1.38E-03**	0.362	**1.02E-03**	-0.114	0.552

Table 8.7. Associations of SNP alternative alleles with phenotypes conditional on the repeat structure (Table 4).

Note rs151182735, rs10168, rs2250063 perfectly correlated with the 3-repeat allele, so these cannot be tested when the repeat structure effect being conditioned on contains that allele. Best-fitting model for DHFR expression contains effects of either rs1105524 or rs1650697 and number of 3-repeat alleles (bold). rs1677658 explains the association between repeat structure and AAO in the combined HD+DM1 sample (bold). Otherwise, the association between SNPs and phenotypes can be explained by the repeat alleles.

The associations of SNPs with phenotypes were conditioned on the effects of *MSH3* repeat alleles (Table 8.7). As rs151182735, rs10168, rs2250063 perfectly correlated with 3a, their independent effects could not be determined (Table 5). With the exception of rs1677658 (LD with 3a: r^2=0.610) and rs1650697 (LD with 3a: r^2=0.143), whose alternative alleles were associated with delayed and early AAO respectively in the combined HD and DM1 meta-analysis (p=0.015 and p=0.029, Table 7), there was no significant evidence for association between SNPs and expansion rate, onset or progression independent of repeat alleles.

Considering variants with minor allele frequency >0.1 and all of the repeat alleles, we observed 25 haplotypes in the region, named Hap1 to Hap25 (Table 8). The 3a repeat allele occurs on both Hap1 and Hap2, which differ only in the presence of the rs1677658 alternative allele on the commoner Hap2. Hap1 was associated with reduced somatic expansion in DM1 (p=0.032) and slower progression in HD (p=0.020), whereas Hap2 was associated with reduced somatic expansion (p=0.021) and delayed onset ($p=4.03\times10^{-5}$) in both HD and DM1, and with slower progression ($p=1.64\times10^{-5}$) and reduced expression of *MSH3* (p=0.024) and *DHFR* ($p=1.12\times10^{-3}$) in HD (Table 9).

Overall, this analysis clarifies the sequence and variants present in *MSH3* exon 1 and demonstrates that *MSH3* repeat variants are associated with disease phenotypes in both HD and DM1.

Haplotype ID	Repeat id	rs151182735	rs10168	rs2250063	rs1105525	rs1105524	rs1650697	rs1677658	Frequency	
									DM1	HD
Hap1	3a	1	1	1	0	1	1	0	0.108	0.081
Hap2	3a	1	1	1	0	1	1	1	0.159	0.171
Hap3	3b	1	1	1	0	1	1	1	0.000	0.002
Hap4	5a	0	0	0	0	0	1	0	0.004	0.007
Hap5	5b	0	0	0	0	1	1	0	0.000	0.002
Hap6	6a	0	0	0	0	0	0	0	0.004	0.000
Hap7	6a	0	0	0	0	0	1	0	0.297	0.331
Hap8	6a	0	0	0	0	1	0	0	0.002	0.000
Hap9	6a	0	0	0	0	1	1	0	0.120	0.083
Hap10	6a	0	0	0	1	1	0	0	0.132	0.120
Hap11	6a	0	0	0	1	1	1	0	0.037	0.044
Hap12	6b	0	0	0	0	1	0	0	0.002	0.007
Hap13	6c	0	0	0	0	1	1	0	0.000	0.002
Hap14	7a	0	0	0	0	1	0	0	0.069	0.093
Hap15	7b	0	0	0	0	0	1	0	0.002	0.002
Hap16	7b	0	0	0	0	1	1	0	0.002	0.005
Hap17	7b	0	0	0	1	1	0	0	0.002	0.000
Hap18	7c	0	0	0	0	1	1	0	0.002	0.002
Hap19	7d	0	0	0	0	1	0	0	0.000	0.002
Hap20	7e	0	0	0	0	0	1	0	0.000	0.002
Hap21	8a	0	0	0	0	0	0	0	0.004	0.000
Hap22	8a	0	0	0	0	1	0	0	0.051	0.039
Hap23	8b	0	0	0	0	1	0	0	0.000	0.002
Hap24	9a	0	0	0	0	0	1	0	0.002	0.000
Hap25	9b	0	0	0	0	1	0	0	0.002	0.000

Table 8.8. MSH3 exon 1 region haplotypes.
For variants, 0 represents a reference allele and 1 represents an alternative allele.

Haplotype ID	Regression with relative rate of somatic expansion								Regression with residual variation in age at onset								Regression with progression score			*MSH3* expression (blood)			*DHFR* expression (blood)			*MSH3* expression (cortex, imputed)			*DHFR* expression (cortex, imputed)		
	HD			DM1			Meta-analysis		HD			DM1			Meta-analysis		HD			HD			HD			HD			HD		
	β	*p*	r2	β	*p*	r2	Z-score	*p*	β	*p*	r2	β	*p*	r2	Z-score	*p*	β	*p*	r2	β	*p*	r2	β	*p*	r2	β	*p*	r2	β	*p*	r2
Hap1	0.045	0.796	-0.004	-0.091	**0.032**	0.015	-1.399	0.162	0.024	0.385	-0.002	-1.596	0.391	-0.001	-0.153	0.878	-0.419	**0.020**	0.021	0.149	0.408	-0.003	-0.208	0.246	0.004	-1.158	**1.92E-11**	0.191	-0.875	**9.32E-07**	0.105
Hap2	-0.321	**6.53E-03**	0.030	-0.021	0.523	-0.002	-2.300	**0.021**	0.061	**3.83E-03**	0.057	4.250	**3.28E-03**	0.034	4.106	**4.03E-05**	-0.521	**1.64E-05**	0.081	-0.285	**0.024**	0.040	-0.403	**1.12E-03**	0.091	-1.347	**1.51E-39**	0.564	-0.033	0.792	-0.004
Hap3																															
Hap4																															
Hap5																															
Hap6																															
Hap7	-0.013	0.885	-0.005	0.069	**0.010**	0.023	1.793	0.073	-0.025	0.106	0.013	-2.357	**0.043**	0.014	-2.589	**0.010**	0.274	**3.83E-03**	0.035	-0.141	0.173	0.009	0.431	**1.20E-05**	0.165	0.251	**8.61E-03**	0.028	0.064	0.512	-0.003
Hap8																															
Hap9	-0.037	0.814	-0.005	0.021	0.598	-0.003	0.228	0.820	0.007	0.795	-0.007	-3.719	**0.039**	0.015	-1.481	0.139	-0.037	0.822	-0.005	-0.279	0.181	0.008	-0.064	0.759	-0.009	0.162	0.330	0.000	0.221	0.187	0.004
Hap10	0.266	0.083	0.010	0.005	0.904	-0.004	1.265	0.206	-0.015	0.527	-0.005	1.142	0.495	-0.002	0.157	0.875	0.283	0.076	0.010	0.251	0.138	0.012	-0.227	0.177	0.008	1.003	**4.91E-11**	0.184	0.503	**1.70E-03**	0.042
Hap11	0.074	0.749	-0.004	-0.094	0.173	0.004	-0.784	0.433	0.013	0.718	-0.007	1.328	0.642	-0.004	0.589	0.556	-0.081	0.735	-0.004	-0.154	0.508	-0.006	0.451	**0.049**	0.028	-0.136	0.575	-0.003	0.684	**4.62E-03**	0.033
Hap12																															
Hap13																															
Hap14	0.112	0.485	-0.002	-0.132	**0.011**	0.022	-1.393	0.164	-0.019	0.519	-0.005	3.983	0.076	0.010	1.017	0.309	-0.046	0.781	-0.004	0.555	**8.55E-04**	0.096	-0.137	0.421	-0.003	0.893	**3.60E-08**	0.132	-0.576	**5.66E-04**	0.051
Hap15																															
Hap16																															
Hap17																															
Hap18																															
Hap19																															
Hap20																															
Hap21																															
Hap22	0.302	0.242	0.002	0.037	0.531	-0.002	1.254	0.210	-0.021	0.615	-0.006	1.596	0.525	-0.003	0.199	0.842	0.705	**7.87E-03**	0.029	0.696	**8.26E-03**	0.058	-0.116	0.663	-0.008	1.119	**2.19E-05**	0.079	0.443	0.102	0.008
Hap23																															
Hap24																															
Hap25																															

Table 8.9. MSH3 exon 1 region haplotypes and their association with phenotypes in DM1 and HD.

An additive genetic model was used to score haplotypes for linear regression analysis. Relative rate of somatic expansion and age at onset were corrected for CAG·CTG length (DM1 and HD) and variant repeats (DM1).

8.5.3 *MSH3* expression in lymphoblasts

The mean expression level of *MSH3* was 36.4% (sem = 10.2%, p = 0.021) lower in lymphoblasts derived from six TRACK-HD 3a repeat homozygotes, relative to six 6a homozygotes, on western blot (Figure 7). The difference was not significant on qPCR, and investigation in a larger sample is warranted (Figure 7).

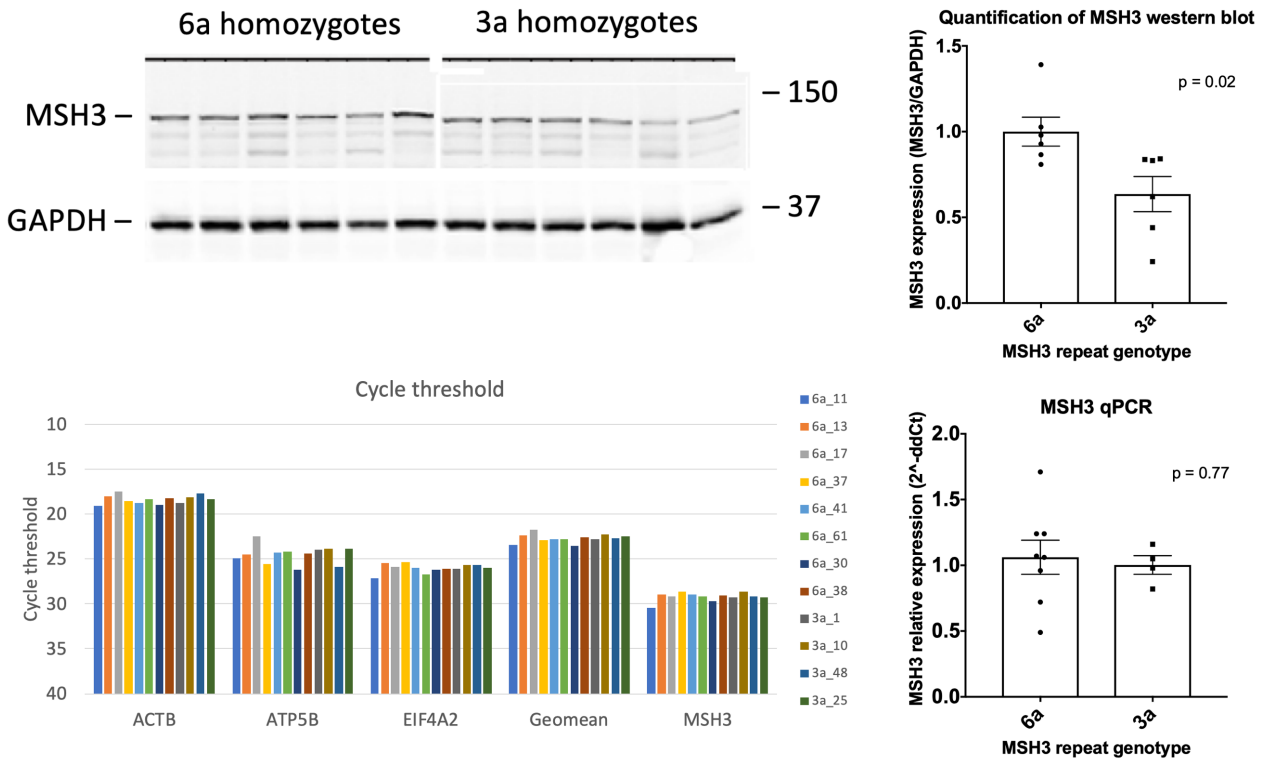


Figure 8.7. MSH3 expression in 6a and 3a repeat homozygotes.
Top left *– western blot for MSH3 and actin.* ***Top right*** *– quantification of western blot. Mean expression level of MSH3 was 36.4% lower in lymphoblasts derived from six TRACK-HD 3a repeat homozygotes, relative to six 6a homozygotes, on western blot (p = 0.021).* ***Bottom left*** *– qPCR of housekeeping genes (ACTB, ATP5B and EIF4A2), Geomean of housekeeping genes, and MSH3 cycle thresholds. Samples are named by repeat genotype, suffixed with subject ID number.* ***Bottom right*** *– MSH3 expression relative to the mean of 6a repeat homozygotes, calculated using the 2^-ddCt method.*

8.5.4 *MSH3* and *DHFR* expression in blood is associated with repeat alleles

Each 3a allele was associated with reduced *DHFR* expression in blood ($p=2.48\times10^{-4}$) and homozygosity for 3a was associated with reduced *MSH3* expression (p=0.0273, Figure 8A), whereas each 7a or 8a allele was associated with increased *MSH3* expression ($p=8.55\times10^{-4}$ and $p=8.26\times10^{-3}$ respectively). The sum of *MSH3* repeat lengths on both alleles appeared to correlate with *MSH3* ($p=7.00\times10^{-3}$) and *DHFR* expression ($p=1.76\times10^{-3}$), which would suggest increasing repeat length increases expression of both (Figure 9). However, a more detailed analysis of *MSH3* repeat alleles (Table 4) shows the number of seven- or eight-repeat alleles is associated with increased expression of *MSH3* ($p=4.53\times10^{-6}$), and that this explains the apparent association with the sum of repeat lengths.

In the detailed analysis, the number of three-repeat alleles was associated with reduced *DHFR* expression (Figure 8B, $p=2.33\times10^{-4}$), and this was sufficient to explain the apparent association of *DHFR* expression with other repeat alleles (Table 4), including that observed with increasing total repeat length. *DHFR* and *MSH3* expression are correlated (Fig. 8C, $r^2=0.120$, $p=2.06\times10^{-4}$). However, association between *DHFR* and three-repeat alleles remains significant after correcting for *MSH3* expression ($p=7.51\times10^{-4}$), and association between *MSH3* and seven- or eight-repeat alleles remains significant

after correcting for *DHFR* expression ($p=1.30\times10^{-7}$). In the best-fitting model for *DHFR* expression, the alternative allele at rs1105524 (LD with 3a: $r^2=0.192$) increases and rs1650697 decreases *DHFR* expression independently of the three-repeat alleles (Table 7). Otherwise, the repeat allele is the major determinant of *MSH3* and *DHFR* expression, and there is no evidence of independent SNP effects.

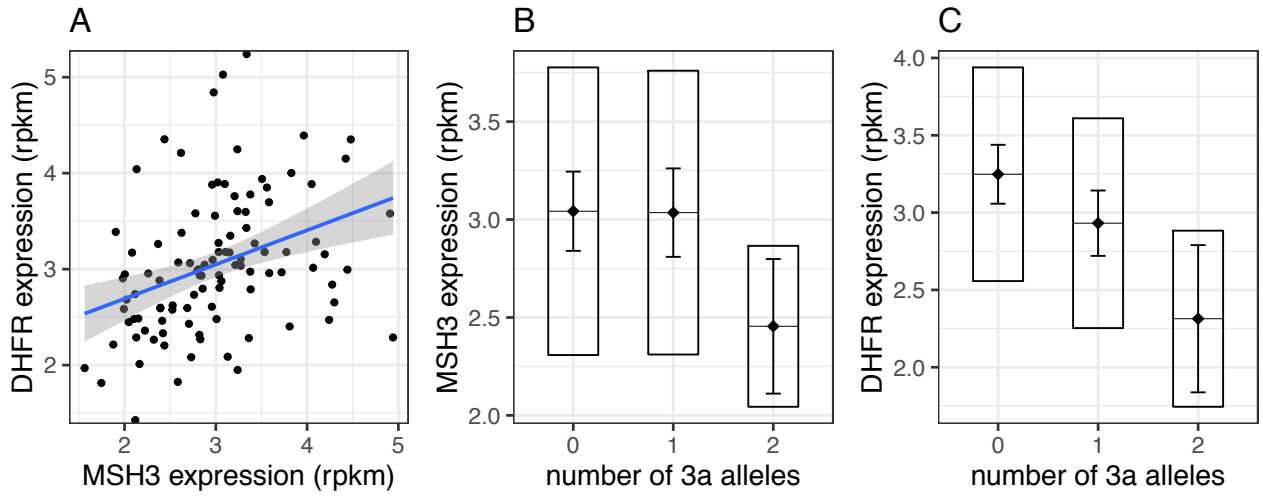


Figure 8.8. Association of the MSH3 3a allele with MSH3 and DHFR expression in HD whole blood.
Whole blood RNA-Seq in a subset of 108 HD subjects. ***(A)*** *Significant correlation between MSH3 and DHFR expression levels (r=0.358, $p=2.06\times10^{-4}$). Grey area around the blue regression line represents 95% confidence interval of the model.* ***(B)*** *Homozygosity for MSH3 3a repeat allele is associated with lower MSH3 expression in blood (p=0.028).* ***(C)*** *MSH3 3a repeat allele is associated with lower DHFR expression (p=2.33×10-4). Rpkm - Reads Per Kilobase of transcript per Million mapped reads. In boxplots, the diamond and horizontal line spanning the diamond indicate the mean, the box indicates the standard deviation and the whiskers indicate the 95% confidence intervals of the mean.*

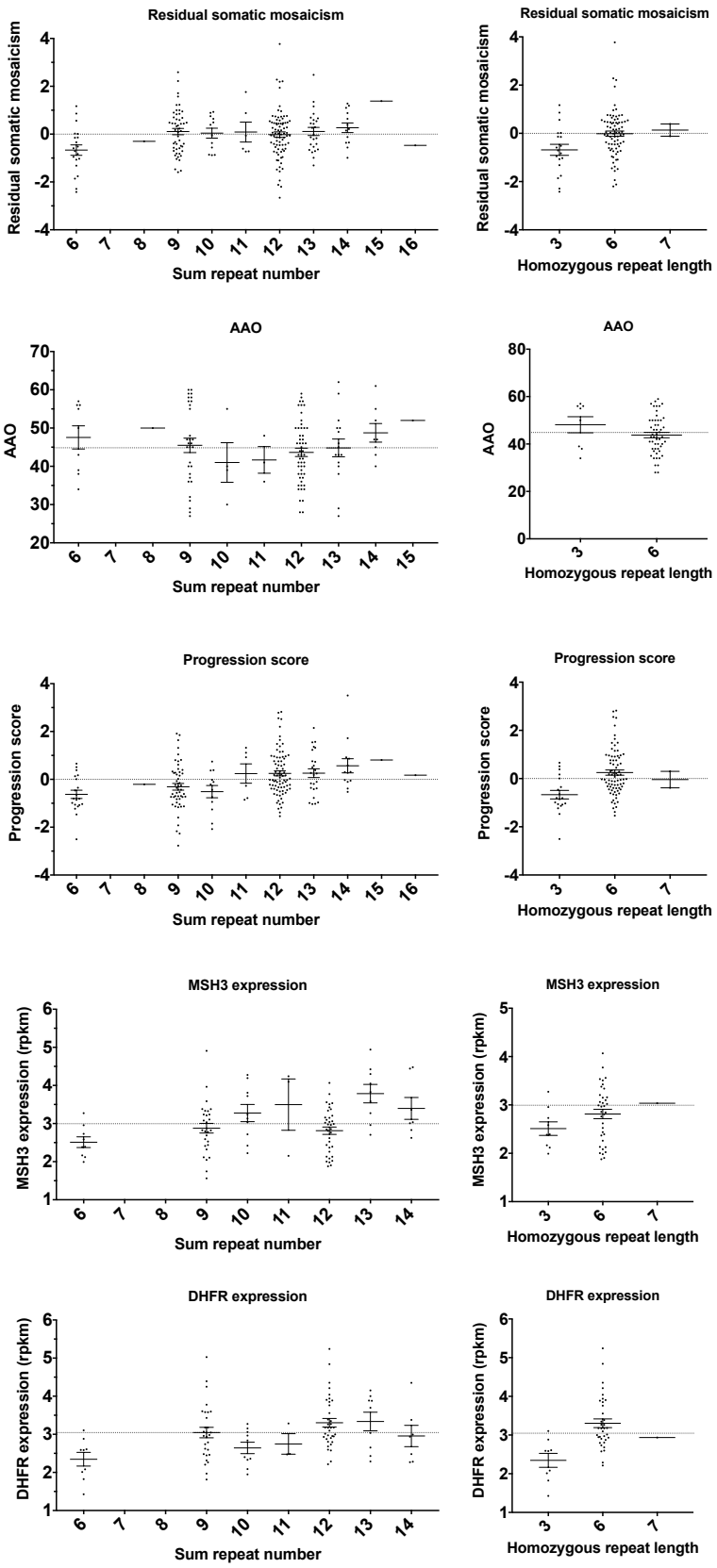


Figure 8.9. MSH3 repeat length correlation with somatic expansion, age at onset, progression score and blood expression of MSH3 and DHFR in HD.

Left column *– sum of the number of MSH3 repeats on both alleles,* ***right column*** *– MSH3 repeat length in repeat homozygotes. Rows in order – residual rate of somatic expansion, age at onset (AAO), progression score, MSH3 and DHFR expression. 79/81 subjects with a sum of 12 repeats are 6 repeat homozygotes.*

8.5.5 *MSH3* expression in cortex is associated with somatic expansion, disease onset and progression in HD

In a TWAS, increased expression of both *MSH3* and *DHFR* in prefrontal cortex (CMC, 2017) was associated with faster progression in TRACK-HD (Hensman Moss et al., 2017b) at similar levels of significance ($p=2.52\times10^{-6}$ and $p=4.08\times10^{-6}$ respectively, see Appendix), making it difficult to distinguish which is more functionally relevant. This ties in with the observation that SNPs significantly associated with somatic expansion, AAO and progression (Table 8.6) were eQTLs for both *MSH3* and *DHFR* in CMC data. Notably, however, increased *MSH3* expression was significantly associated with early onset ($p=1.71\times10^{-3}$) in a TWAS of the GeM dataset (GeM-HD, 2015), while *DHFR* expression was not significantly associated with onset (Appendix). This favours *MSH3* over *DHFR* expression as a modifier of HD disease course.

In the TRACK-HD cohort, 3a genotype correlates with *MSH3* expression in cortex (p=4.55E-75, Table 8.3). It's correlation with *DHFR* expression is not as strong (p=2.49E-03, Table 8.3)

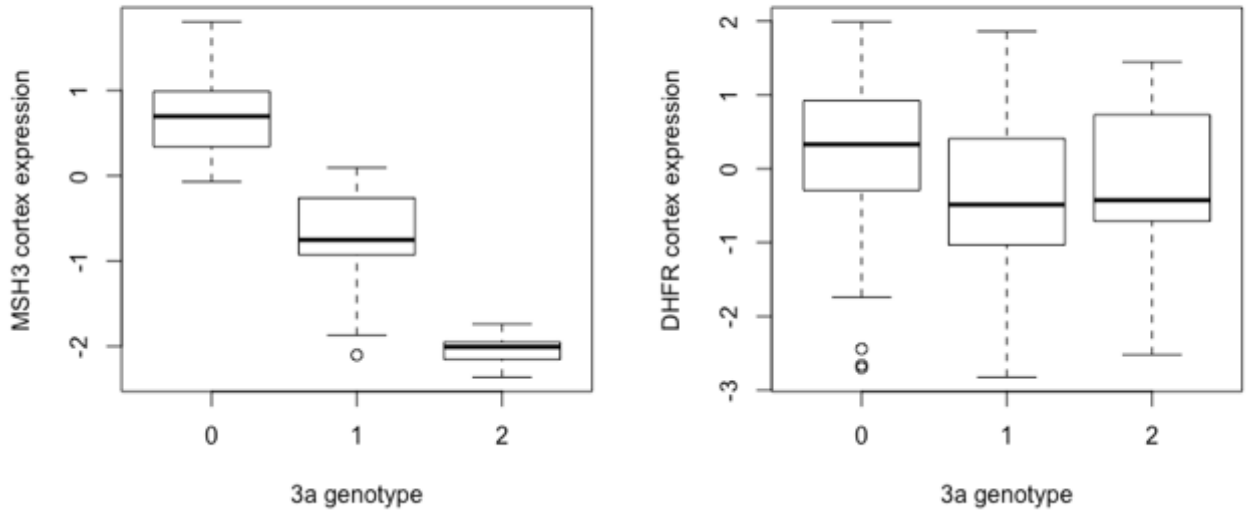


Figure 8.10. Association of the MSH3 3a allele with MSH3 and DHFR expression in the TRACK-HD prefrontal cortex TWAS.
The MSH3 3a repeat allele is associated with lower MSH3 expression in prefrontal cortex (p=4.55E-75). It is less significantly correlated with reduced DHFR expression in cortex (p=2.49E-03).

The sum of each subject's *MSH3* repeat lengths correlated with *MSH3* expression in cortex (p≤2.2e-16) more strongly than in blood (p=0.007002). It also correlated more strongly than with *DHFR* expression in cortex (p=0.009946). This suggests that increasing *MSH3* repeat length increases *MSH3* expression in the central nervous system.

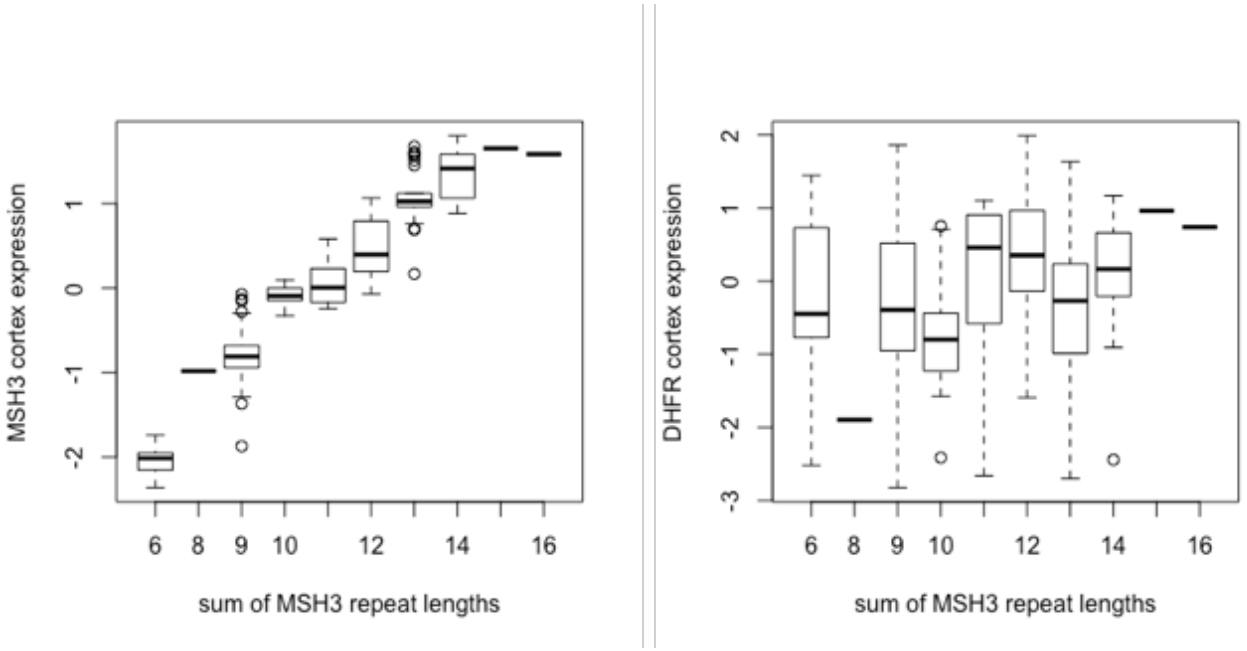


Figure 8.11. MSH3 repeat length correlation with prefrontal cortex expression of MSH3 and DHFR in TRACK-HD.
Left– *Sum of MSH3 repeat length is more strongly correlated with MSH3 expression in cortex (p=< 2.2e-16) than blood (p=0.007002).* ***Right*** *MSH3 repeat length correlation with DHFR expression in cortex is not as strong as for MSH3 (p=0.009946).*

In the TRACK-HD TWAS, prefrontal cortex *MSH3* expression was associated with CAG expansion rate in blood (p=0.0143). This suggests that variants which predict *MSH3* expression in cortex are also associated with CAG expansion rate. Prefrontal cortex *DHFR* expression was not associated with CAG expansion rate, favouring *MSH3* as the driver of somatic expansion.

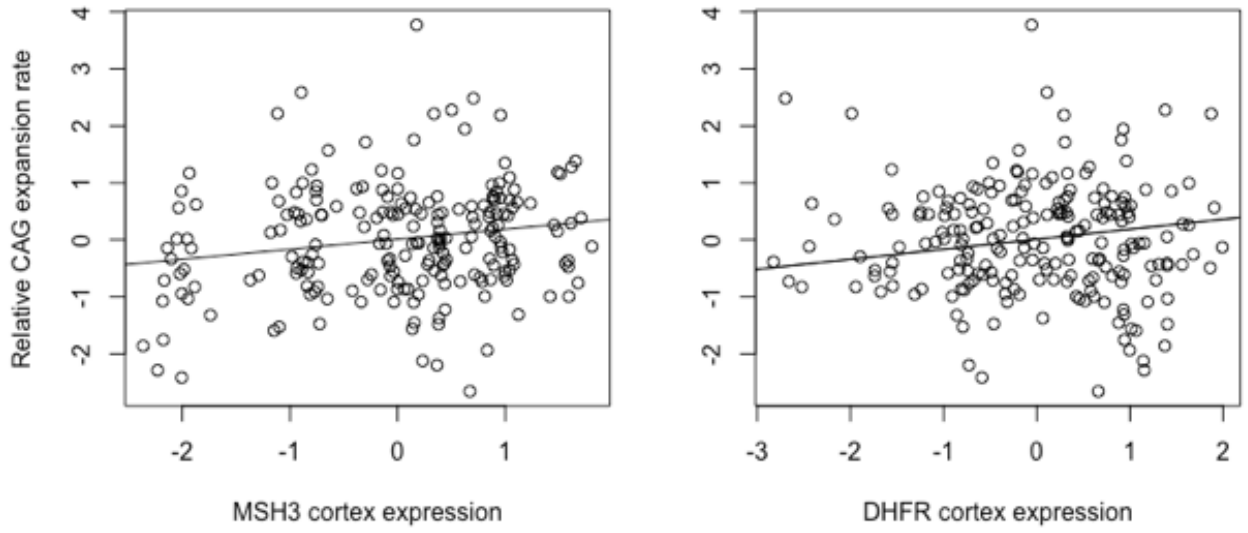


Figure 8.12. CAG repeat expansion correlation with MSH3 or DHFR expression in TRACK-HD prefrontal cortex.
Left *– Blood CAG expansion rate correlates better with MSH3 expression level in cortex (p=0.0143) than expression level blood (p=0.6248).* ***Right*** *– Blood CAG expansion rate was not significantly associated with DHFR expression in prefrontal cortex (p=0.6054).*

8.6 Discussion

8.6.1 *MSH3* repeat alleles

MSH3 has recently been identified as a genetic modifier of somatic instability in DM1 (Morales et al., 2016), and progression in HD (Hensman Moss et al., 2017b). The *MSH3* signal in the GWAS of HD progression was driven by an imputed SNP, rs557874766, located within a 9 bp tandem repeat sequence in exon 1 of *MSH3*, which is also in the 5' UTR of *DHFR* on the opposite strand. *MSH3* and *DHFR* are organised head-to-head, transcribed in opposite directions and are regulated by the same promoter (Drummond, 1999). This chapter demonstrates that rs557874766 is an alignment artefact and corresponds to a three-repeat allele, 3a. 3a was the shortest repeat allele observed and is likely ancestral. At the protein level, *in silico* modelling predicts that 3a results in the loss of a surface α-helix (Kallberg et al., 2012) at the N-terminus of MSH3.

A total of 16 *MSH3* repeat alleles were observed, varying in sequence and length from three to nine repeats. 6a and 3a are the first and second commonest in this European cohort, though previous studies suggest a seven-repeat allele may be second commonest in East Asian populations (Nakajima et al., 1995). In HD, 3a was associated with reduced somatic expansion, delayed onset and slower progression. In DM1, each 3a allele showed a trend towards reduced somatic expansion and delayed onset but was significant with both measures in meta-analysis of HD and DM1. Longer seven-repeat alleles were associated with reduced somatic expansion only in DM1. Whether this reflects a subtle difference in MSH3 biology between the two disorders, or simply a sampling error, remains to be determined.

The *MSH3* repeat lies within a region of basic amino acids 23 residues downstream of the PCNA interaction domain (PIP box) (Kleczkowska et al., 2001, Clark et al., 2000, Flores-Rozas et al., 2000, Finn et al., 2016), 7 residues upstream of the EXO1 binding region (Schmutte et al., 2001) and 58 upstream of the MSH2 binding region (Dragileva et al., 2009, Tome et al., 2013a, Campregher et al., 2012), all of which are involved in mismatch repair. PCNA is a sliding clamp that participates in DNA replication, but in MMR it delivers MSH proteins to mismatches and increases mismatch binding specificity (Flores-Rozas et al., 2000). Exonuclease 1 (EXO1) excises the daughter strand after mismatch recognition, as well as being involved in end resection during homologous recombination. Whilst it is involved in MMR, it is not absolutely required, with MutL or other nucleases able to compensate (Goellner et al., 2015). Sequence conservation suggests this region may form a short, flexible connector domain involved in protein interaction (Kleczkowska et al., 2001). The *MSH3* repeat region is poorly conserved between species, with other mammals having between zero and five repeats. This lack of evolutionary constraint suggests functional redundancy in the MMR pathway and a lack of a major effect of N-terminal MSH3 variation outside the context of repeat expansion disease. Unlike other MMR components, germline heterozygous *MSH3* mutations are not associated with increased risk of cancer, most likely because MSH2/MSH6 can also initiate repair at replication errors (Haugen et al., 2008, Edelmann et al., 2000, Jiricny, 2006).

8.6.2 Repeat-flanking variants

Three non-coding variants 5' of the repeat were in near complete LD with 3a, so it is not possible to determine their independent effects on disease phenotypes. All three are associated with reduced *MSH3* expression in multiple tissues, including cortex (CMC and GTEx). Controlling for repeat alleles, no SNPs were significantly associated with phenotypes, except the intronic rs1677658 and the exon 1 rs1650697 variants, which contributed to delayed or early onset

respectively in the combined HD and DM1 dataset. rs1677658 was associated with reduced *MSH3* and *DHFR* expression (CMC and GTEx), whereas rs1650697 was associated with increased *DHFR* in HD blood, as well as multiple tissues in GTEx.

Hap2, the *MSH3* haplotype most significantly linked with reduced somatic expansion and delayed onset in HD and DM1, and with slower progression in HD, contains the 3a allele, along with alternative alleles of non-coding variants rs151182735, rs10168, rs2250063, which are in complete LD with it, and rs1677658. It is thus difficult to assess which (if any) *MSH3* variants (repeats or SNPs) are driving associations with disease phenotypes, and further investigation in a larger sample is warranted.

8.6.3 Transcriptomic analysis

Whole blood transcriptomic analysis in a subset of the HD patients found the 3a allele was associated with reduced expression of *MSH3* and *DHFR,* and seven- or eight-repeat alleles with increased *MSH3* expression. *DHFR*, which shares a promoter with *MSH3* (Drummond, 1999), is a ubiquitously expressed enzyme involved in purine, thymidylic acid and amino acid synthesis, but has not previously been implicated in HD pathogenesis.

The TWAS found that increased expression of *MSH3* and *DHFR* in cortex are associated with faster HD progression (Hensman Moss et al., 2017b). While *MSH3* expression was significantly associated with early onset in the GeM TWAS ($p=1.71\times10^{-3}$) (GeM-HD, 2015), *DHFR* expression was not associated with disease course. This is consistent with HD mouse brain, in which expression of *MSH3*, but not *DHFR*, correlates with somatic expansion (Tome et al., 2013a). The 3a repeat was associated with reduced *MSH3* expression in cortex (p=4.55E-75), more so than *DHFR* (p=2.49E-03, Fig. 8.10), and *MSH3* expression level in cortex was associated with CAG repeat expansion rate (p=0.0143, Fig. 8.12). This suggests variants that predict *MSH3* expression in cortex are also associated with CAG expansion. HD TWA studies also found that increased *FAN1* expression in cortex was associated with delayed onset in GeM-HD ($p=2.80\times10^{-12}$) (GeM-HD, 2015) and slower progression in TRACK-HD and REGISTRY ($p=1.58\times10^{-4}$) (Hensman Moss et al., 2017b). This suggests that *MSH3* expression is deleterious and *FAN1* is protective in HD brain.

8.6.4 *MSH3* variants modify somatic instability and disease severity in HD and DM1

Collectively, these results suggest the *MSH3* 3a repeat allele reduces somatic expansion and improves phenotype in both HD and DM1, potentially through altering *MSH3* expression levels. However, given the proximity of the repeat region to MMR protein binding domains, the 3a allele could also alter MSH3 function in the recognition and repair of insertion-deletion loops, double-strand breaks or single-strand annealing (Lyndaker and Alani, 2009, Schmidt and Pearson, 2016). Repetitive DNA sequences form unusual secondary structures such as slipped strands, hairpin loops, G-quadruplexes and R-loops (Mirkin, 2007, Neil et al., 2017), the stability of which correlates with expansion (Gacy et al., 1995). MSH3 may recognise these structures (Owen et al., 2005) and initiate repair, during which out of register synthesis could result in repeat expansion (Neil et al., 2017, Khan et al., 2015). Together, these results suggest a common mechanism, involving somatic expansion, operates *in vivo* in distinct trinucleotide repeat diseases to influence disease course. Therefore, modulation of MSH3 has significant therapeutic potential in a range of diseases caused by repeat expansions.

8.7 Summary

The mismatch repair gene *MSH3* has been implicated as a genetic modifier of the CAG·CTG repeat expansion disorders Huntington's disease (HD) and myotonic dystrophy type 1 (DM1). A recent HD genome-wide association study found

rs557874766, an imputed single nucleotide polymorphism (SNP) located within a polymorphic 9 bp tandem repeat in *MSH3*/*DHFR*, as the variant most significantly associated with progression in HD. Using Illumina sequencing in HD and DM1 subjects, this chapter shows that rs557874766 is an alignment artefact, the minor allele for which corresponds to a three-repeat allele in *MSH3* exon 1 that is associated with a reduced rate of somatic CAG·CTG expansion (p=0.004) and delayed disease onset (p=0.003) in both HD and DM1, and slower progression ($p=3.86\times10^{-7}$) in HD. RNA-Seq of whole blood in the HD subjects found that repeat variants are associated with *MSH3* and *DHFR* expression. A transcriptome-wide association study in the HD cohort found increased *MSH3* and *DHFR* expression are associated with disease progression, and increased *MSH3* expression in cortex was associated with increased somatic expansion. These results suggest that variation in the *MSH3* exon 1 repeat region influences somatic expansion and disease phenotype in HD and DM1, and suggests a common DNA repair mechanism operates in both repeat expansion diseases.

8.8 Publications related to this chapter

The work presented in this chapter was published in:

MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. **Flower M.***, Lomeikaite V.*, Ciosi M., Cumming S., Morales F., Lo K., Hensman Moss D., Jones L., Holmans P., the TRACK-HD Investigators, the OPTIMISTIC Consortium, Monckton D.G[#]. and Tabrizi S.J.[#] ***Brain*** (accepted March 2019).

* These authors should be regarded as joint first authors.

[#] These authors jointly supervised the work.

Chapter 9 Conclusions and future work

9.1 Conclusions

9.1.1 DNA repeat instability in disease

Expansion of repeated DNA sequences cause dozens of human diseases, and though the repeats occur in different proteins and genomic contexts, and have varied pathogenic mechanisms ranging from silencing expression (Colak et al., 2014) and RNA foci (Thornton, 2014) to RAN translation (Zu et al., 2011, Banez-Coronel et al., 2015, Cleary and Ranum, 2014) and aggregate formation (Ross and Tabrizi, 2011), they usually have a neurological phenotype, suggesting the nervous system is particularly susceptible to this type of mutation (Neil et al., 2017). The repeats are inherently unstable and tend to expand over time in the tissues most vulnerable in each disease (Thornton et al., 1994, Zatz et al., 1995, Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009, Jedele et al., 1998, Tanaka et al., 1999). Expansion correlates with disease onset, progression and severity in animal models and patients (Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009, Gonitel et al., 2008, Lee et al., 2011a, Mangiarini et al., 1997, Ashizawa et al., 1993, Lia et al., 1998, Fortune et al., 2000, Kennedy and Shelbourne, 2000). Instability likely happens during DNA repair or transcription, rather than during replication, as it continues when the cell cycle is arrested (Gomes-Pereira et al., 2014b), is negatively correlated with cell cycle pathways (Lee et al., 2010), and occurs in the post-mitotic neurons of animal models and patients (Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009, Gomes-Pereira et al., 2014b). Interruptions in the repeat sequence, which reduce formation of abnormal structures such as hairpin loops, protect against instability (Massey and Jones, 2018, Menon et al., 2013, Pearson et al., 1998, Sobczak and Krzyzosiak, 2004, Kraus-Perrotta and Lagalwar, 2016). This suggests the DNA secondary structure is important in modulating susceptibility to instability.

9.1.2 CAG repeat expansion *in vitro*

Whilst CTG (Ashizawa et al., 1996, Bidichandani et al., 1999, Ashizawa et al., 1993, De Temmerman et al., 2008, Seriola et al., 2011a, Du et al., 2013a), GAA (Lai et al., 2014, Ku et al., 2010, Du et al., 2012a) and ATTCT (Lin and Ashizawa, 2003, Liu et al., 2007) repeats appear readily unstable in cultured cells, no cell models of robust and significant CAG repeat expansion exist (Kovtun et al., 2007, Jonson et al., 2013a, Jacquet et al., 2015, Mollica et al., 2016). In chapters 5 and 6, several were developed, including human osteosarcoma and neural stem cells transduced with *HTT* exon 1 containing 97 or 129 CAG repeats, and patient-derived lymphoblastoid (LB), stem cell and differentiated medium spiny neurons (MSN) from HD subjects with 109 and 125 CAG repeats. This is the first demonstration of CAG repeat instability in cultured patient MSNs, the cell type most vulnerable in HD and the most physiologically relevant cell model in which to study HD pathogenesis. Expansion was repeat length-dependent, with rate increasing exponentially after a trigger point between 73 and 97 CAG repeats.

Human cell lines and patient-derived LBs, stem cells and differentiated MSNs with up to 73 CAG repeats were stable in long term culture up to 29 weeks, even on exposure to chronic oxidative stress. This is consistent with previous studies which have shown that to trigger expansion, cultured cells must harbour over 64 repeats (Cannella et al., 2009). Human disease is most commonly caused by repeat lengths around 40 CAG, and expansion likely occurs slowly over decades until a toxic threshold is reached, after which time the polyglutamine tract confers toxicity in vulnerable striatal neurons,

causing clinical onset (Kaplan et al., 2007, Kennedy et al., 2003, Swami et al., 2009). The lack of observable instability may reflect the rarity of expansion events at these repeat lengths, and extremely long-term culture may be required to observe a change.

9.1.3 DNA repair modifies the course of repeat expansion diseases

In HD, repeat length is the main influence on disease course, though a significant proportion of the variation in disease onset is due to variation elsewhere in the genome (Gusella et al., 2014, Wexler et al., 2004a). GWAS studies probing this variation have identified loci on chromosomes 15, 8, 5 and 3 that modify onset or progression (GeM-HD, 2015, Lee et al., 2017, Hensman Moss et al., 2017b). These are likely underlain by *FAN1*, *RRM2B*, *MSH3* and *MLH1* respectively, suggesting that, after repeat length, DNA repair is the strongest determinant of disease course. Supporting this, pathway analyses in each GWAS highlighted sets of DNA repair genes, particularly mismatch repair, and *MSH3* variants have been shown to modify expansion of CTG repeats in myotonic dystrophy type 1 (DM1) patients (Morales et al., 2016). Chapter 3 showed that DNA repair variants from HD GWA studies also influence onset in the other polyglutamine diseases (Bettencourt et al., 2016), suggesting a common mechanism operates in conditions caused by repeat expansion. The lead SNP, rs3512 in *FAN1*, was associated with 1.3 year delayed onset in the GeM-HD GWAS (p = 5.28E-13), but variants in *PMS2*, a component of the MutLα MMR complex, and *RRM2B*, involved in nucleotide synthesis, were also significant.

9.1.4 DNA repair drives repeat instability

The DNA damage response (DDR) is a series of overlapping pathways that sense and repair lesions that occur continually throughout the body (Ciccia and Elledge, 2010). The nervous system appears particularly susceptible to DNA damage, with DDR defects implicated in many neurological diseases, including oxidative lesions in Alzheimer's, Parkinson's and amyotrophic lateral sclerosis (ALS) (McKinnon, 2009, Canugovi et al., 2013) and strand breaks in several hereditary ataxias (Jackson, 2002, Madabhushi et al., 2014, Suberbielle et al., 2013, El-Khamisy et al., 2005, Clements et al., 2004).

In repeat expansion diseases, however, it appears the DDR actively promotes instability. Several pathways have been implicated, but mismatch repair (MMR) is the strongest driver (Castel et al., 2010, Slean et al., 2008), with depletion of complexes MutSβ (MSH2/MSH3), MutLα (MLH1/PMS2) and MutLγ (MLH1/MLH3) protecting against repeat expansion in cell and animal models of HD (Lopez Castel et al., 2010, Manley et al., 1999, Dragileva et al., 2009, Tome et al., 2013a, Gomes-Pereira, 2004, Gomes-Pereira et al., 2014b, Pinto et al., 2013b, Pinto et al., 2013a), DM1 (Nakatani et al., 2015b, Stevens et al., 2013, Du et al., 2013b, Seriola et al., 2011b, Nakatani et al., 2015c, Williams and Surtees, 2015, Kantartzis et al., 2012, Dragileva et al., 2009, Tome et al., 2013a, van den Broek et al., 2002, Foiry et al., 2006, Morales et al., 2016), fragile X (Lokanga et al., 2014) and Friedreich's ataxia (Bourn et al., 2012, Zhao et al., 2015b, Ezzatizadeh et al., 2012), as well as in HD (Hensman Moss et al., 2017b) and DM1 patients (Morales et al., 2016). *Fan1* knockout in a fragile X mouse model has also been shown to accelerate CGG repeat expansion (Zhao and Usdin, 2018). A great challenge currently facing the field is understanding how a system that normally guards genomic stability is instead contributing to cell dysfunction and death (Jiricny, 2006). MutSβ, which normally recognises short insertion-deletion loops (IDL), may bind abnormal DNA secondary structures formed on lagging strand templates (Freudenreich et al., 1997, Kang et al., 1995, Liu et al., 2010a, Panigrahi et al., 2002, Cleary et al., 2010, Nenguke et al., 2003) by repeat sequences (Mirkin, 2007, Neil et al., 2017, McMurray, 2010, Gacy et al., 1995), initiating inaccurate repair that leads to expansion (Schmidt and Pearson,

2016, Williams and Surtees, 2015). The prevalence (Axford et al., 2013) and stability (Gacy et al., 1995) of such structures has been shown to correlate with expansion.

9.1.5 FAN1

9.1.5.1 What is already known about FAN1

FAN1 is a structure-specific endo/exonuclease involved in DNA interstrand crosslink (ICL) repair (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Smogorzewska et al., 2010a) and the recovery of stalled replication forks (Lachaud et al., 2016a, Chaudhury et al., 2014), though its precise role in both processes is unknown (Thongthip et al., 2016, Lachaud et al., 2016a, Lachaud et al., 2016b). It stabilises CGG repeats (Zhao and Usdin, 2018) and its knockout prevents the resolution of double strand breaks (DSB) induced during ICL repair (Thongthip et al., 2016, Lachaud et al., 2016a, Lachaud et al., 2016b). Its PCNA interaction domain is required for recruitment to stalled replication forks (Porro et al., 2017), the UBZ domain is needed for an interaction with FANCD2 that regulates its activity at stalled forks (Chaudhury et al., 2014, Chen et al., 2015, Schlacher et al., 2012, Lachaud et al., 2016a), and its nuclease domain is required for all its currently known functions (Ray Chaudhuri et al., 2012, Ge and Blow, 2010, Lachaud et al., 2016a). Mutations are linked with a recessive renal syndrome with polyploidy (Zhou et al., 2012, Lachaud et al., 2016b, Thongthip et al., 2016), as well as colorectal (Segui et al., 2015b) and pancreatic (Smith et al., 2016) cancers, but unlike other ICL repair genes, they do not cause Fanconi anaemia. FAN1 is known to interact directly with the ID complex (FANCD2 and FANCI), but a substantial proportion of cellular FAN1 binds MLH1 of the MMR MutLα complex (MacKay et al., 2010b). A role has not yet been shown for the latter, but one can speculate FAN1 and MMR interact in a pathway promoting repeat instability, with FAN1 potentially sequestering MLH1 away from the expansion-inducing MutSβ complex.

9.1.5.2 FAN1 depletion sensitises cells to interstrand crosslinks

FAN1 knockout, knockdown or inactivation of its nuclease domain by the p.D960A mutation in human cell lines prevented the resolution of DSBs during ICL repair and sensitised them to chemically-induced ICLs, supporting the known involvement of FAN1 nuclease in the ICL repair pathway (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Smogorzewska et al., 2010a). There was no effect on mismatch or strand break repair, suggesting FAN1 does not participate in these functions. In patient-derived and artificial cell lines expressing *FAN1* variants associated with fast HD progression, including the p.R507H DNA binding domain variant associated with 6 year early HD onset ($p = 9.34E-18$) (GeM-HD, 2015), ICL sensitivity was unaffected. This suggests FAN1 activity at the *HTT* CAG repeat is independent of its ICL repair function.

9.1.5.3 FAN1 protects against CAG repeat expansion

Knockout or knockdown of *FAN1* in human cells expressing a 118 CAG tract, patient-derived iPSCs and post-mitotic MSNs, nearly doubled the expansion rate from around 15-18 days/Q to just 10 days/Q (Chapters 5 and 6). Expansion was FAN1 concentration-dependent, with higher levels increasing the stability of the CAG repeat. Supporting this, the transcriptome-wide association study (TWAS) in chapter 10 showed increased *FAN1* expression was associated with delayed onset and slow progression in HD. Collectively these results suggest the genome-wide chromosome 15 signal linked with HD disease course (GeM-HD, 2015, Lee et al., 2017, Hensman Moss et al., 2017b) results from FAN1 protectively stabilising the CAG tract in a mechanism that is independent of DNA replication (Gomes-Pereira et al., 2014b, Lee et al., 2010, Kennedy et al., 2003, Shelbourne et al., 2007b, Swami et al., 2009). Consistent with this and DM1 cell

models (Gomes-Pereira et al., 2014a, Gomes-Pereira et al., 2001), there was no correlation between mitotic rate and CAG expansion (Chapter 5).

In the R6/2 mouse model of HD, *Fan1* was expressed at relatively high levels in somatically stable tissues such as cerebellum, and low levels in those showing expansion, such as liver. Unfortunately, intrastriatal and peritoneal shRNA-mediated *Fan1* knockdown failed to demonstrate an effect on CAG repeat expansion rate, though this was likely due to relatively poor knockdown by only 23%. In iPSC models, where *FAN1* depletion accelerated expansion rate, 60% knockdown was achieved.

Surprisingly, the nuclease inactivating p.D960A mutation did not affect expansion rate. As FAN1 activity in ICL repair and replication fork recovery require nuclease activity, this suggests a novel function underlies CAG repeat stabilisation. The HD onset-associated p.R507H DNA binding domain variant (GeM-HD, 2015) did not significantly alter expansion rate in the U20S system, though high expression, which as shown above stabilises the repeat, may have obscured a subtle effect on FAN1 function.

9.1.5.4 FAN1 DNA binding

FAN1 has been shown to bind artificial branched DNA substrates (Kratz et al., 2010a, Liu et al., 2010b, MacKay et al., 2010b, Pennell et al., 2014), but its DNA binding preferences in cells have not yet been demonstrated. Through ChIP-qPCR in human cells expressing a 129 CAG repeat and patient-derived LB and iPSCs, a novel interaction was found between FAN1 and CAG repeat DNA in *HTT*, as well as other polyglutamine disease genes. FAN1 also bound other DNA regions, suggesting it does not preferentially interact with CAG repeats. Once again, the p.R507H variant did not alter DNA binding, suggesting it does not affect FAN1 substrate preference.

9.1.6 MSH3

9.1.6.1 What is already known about MSH3

MSH3 heterodimerises with MSH2 to form the MutSβ complex that recognises IDLs and initiates MMR (Tome et al., 2013a, Gonitel et al., 2008). As above, it has been shown to be required for repeat expansion in numerous diseases, including HD. Unlike other MMR proteins, its depletion does not cause cancer, likely because MutSα is able to partially compensate (Edelmann et al., 2000, Jiricny, 2006).

9.1.6.2 MSH3 modifies somatic instability and disease severity in Huntington's disease and myotonic dystrophy type 1

The lead variant in a recent GWAS linking *MSH3* with HD progression was the imputed SNP rs557874766 (Hensman Moss et al., 2017b). In chapter 8, Illumina sequencing of this 9 bp tandem repeat region showed this variant to be an alignment artefact, and that subjects instead had a 3-repeat variant, named 3a, which was associated with reduced blood *MSH3* and *DHFR* expression, reduced somatic expansion in blood, delayed onset and slower progression in HD and myotonic dystrophy type 1 (DM1). Decreased *MSH3* expression has already been shown to reduce repeat expansion and improve disease phenotype in cell and animal models (Dragileva et al., 2009, Tome et al., 2013a, Nakatani et al., 2015b, Stevens et al., 2013, Du et al., 2013b, Seriola et al., 2011b, Nakatani et al., 2015c, Williams and Surtees, 2015, Kantartzis et al., 2012, van den Broek et al., 2002, Foiry et al., 2006, Morales et al., 2016). 16 repeat alleles were observed, with between three and nine repeats, and the relatively common, longer 7 or 8 repeat alleles were associated with increased *MSH3* expression. The repeat region lies between binding sites for MMR proteins PCNA, EXO1 and MSH2 (Kleczkowska et al.,

2001, Clark et al., 2000, Flores-Rozas et al., 2000, Finn et al., 2016, Schmutte et al., 2001, Tome et al., 2013a), and is poorly conserved between species.

Three non-coding variants were in complete linkage disequilibrium (LD) with 3a. The intronic rs1677658 and the exon 1 rs1650697 variants contributed to delayed or early onset respectively. The haplotype most significantly associated with reduced somatic expansion, delayed onset and slower progression in HD and DM1, named Hap2, contained the 3a repeat allele, the non-coding variants rs151182735, rs10168, rs2250063 which are in complete LD with it, and the exon 1 variant rs1677658 (LD with 3a; r^2 = 0.610). Further investigation in a larger sample is warranted to clarify which variants are driving association with disease phenotypes.

A transcriptome-wide association study (TWAS) showed increased *MSH3* expression and reduced *FAN1* expression in prefrontal cortex were associated with early onset and faster progression in HD, suggesting MSH3 is deleterious and FAN1 is protective in the context of a *HTT* CAG expansion. Increased *MSH3* expression in prefrontal cortex was associated with increased somatic expansion.

Genetic variation in *MSH3* may influence disease course through altering *MSH3* expression levels or the protein's interaction with other MMR components. These results suggest a common mechanism involving somatic expansion operates *in vivo* in several repeat expansion disorders, and that modulation of MSH3 has significant therapeutic potential in a range of diseases.

9.1.7 Transcriptional dysregulation in Huntington's disease blood

HD is a systemic condition and the expanded CAG repeat is expressed in mutant HTT (mHTT) throughout the body, causing immune dysfunction (Tai et al., 2007a, Bjorkqvist et al., 2008, Kwan et al., 2012c, Träger et al., 2015), metabolic disruption with weight loss (Carroll et al., 2015), muscle wasting (Busse et al., 2008) and cardiac dysfunction (Lanska et al., 1988, Mihm et al., 2007, Pattison et al., 2008), endocrine disturbance (Saleh et al., 2009) and liver impairment (Carroll et al., 2015) in animal models (Orth et al., 2003) and patients (Turner et al., 2007). In chapter 4, RNA-Seq of whole blood from Track-HD patients identified immune upregulation and dysregulation of DNA repair, RNA processing, energy metabolism. The transcriptional changes in blood correlated with disease severity and replicated the signatures of HD patient caudate (Neueder and Bates, 2014, Hodges et al., 2006) and prefrontal cortex (Labadorf et al., 2015), as well as those of other studies on whole blood (Mina et al., 2016) and hyperactive monocytes (Miller et al., 2016), and significantly overlapped with immune upregulation found in Alzheimer's disease brain (International Genomics of Alzheimer's Disease, 2015, Zhang et al., 2013). These results suggest a key role for the immune system in neurodegenerative disease and support the study of peripheral tissue which, unlike the nervous system, can be sampled minimally invasively and inexpensively from living patients throughout the course of the disease.

9.1.8 Summary

A model is emerging, in which a FAN1/MLH1 complex may bind abnormal secondary structures which are formed by CAG repeat DNA and occur more readily as repeat length increases. Several mechanisms can be hypothesised; **firstly**, the complex may block access of MutSβ (MSH2/MSH3), which would otherwise misidentify these structures as insertion-deletion loops, initiating mismatch repair that erroneously introduces extra repeats. **Secondly**, as MutLα (MLH1/PMS2) independently binds FAN1 and MutSβ (MSH2/MSH3) (MacKay et al., 2010b), FAN1 could sequester MLH1 that would

otherwise act with MSH3 to promote repeat expansion. **Finally**, FAN1 could act on the CAG repeat, promoting accurate repair either directly or as a scaffold for a repair complex. FAN1 has the potential DNA binding and nuclease activity to act directly in repair, though the protection offered against expansion by the nuclease dead p.D960A mutant argues against this.

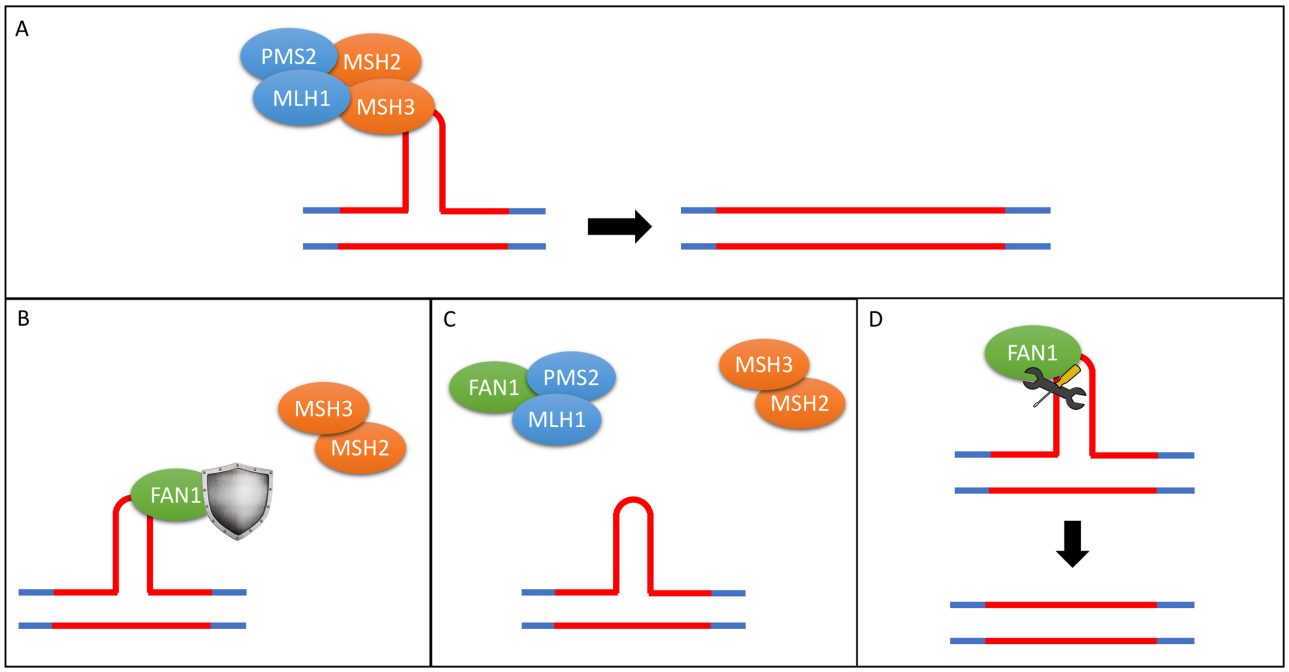


Figure 9.1. Potential mechanisms by which FAN1 may protect against CAG repeat expansion.
A) *MutSβ (MSH2/MSH3, orange) and MutLα (MLH1/PMS2, blue) misidentify abnormal secondary structures formed by CAG repeat DNA (red), such as hairpins, invoking mismatch repair, during which out of register alignment introduces repeat expansion.* ***B)*** *FAN1 (green) may bind CAG repeat DNA, prohibiting access of MutSβ.* ***C)*** *FAN1 may sequester MutLα (MLH1/PMS2) away from MutSβ, preventing MMR.* ***D)*** *FAN1 may act directly at the CAG repeat, promoting accurate repair.*

Neurons appear particularly susceptible to repeat expansion and the dysfunctional proteins it produces, though the reasons for this remain unclear. Therapies increasing FAN1 or reducing MSH3 are expected to restrict CAG repeat expansion, delay onset and slow progression in numerous conditions caused by repeat tracts, including the polyglutamine diseases and myotonic dystrophy, without increasing cancer risk. However, preclinical trials should observe for toxicity associated with FAN1 overexpression and microsatellite instability associated with reduced MSH3 (Haugen et al., 2008).

9.2 Future work

9.2.1 Elucidating the somatic instability network (SIN)

FAN1 and MSH3 are at the core of a wider DDR network involved in CAG repeat instability (Muro et al., 2015), which may include the MMR nuclease EXO1 (McMurray, 2010, Usdin et al., 2015), the DNA clamp PCNA (Usdin et al., 2015, Lopez Castel et al., 2009, Ren et al., 2017), the DNA ligase LIG1 (Usdin et al., 2015, Lopez Castel et al., 2009), the MSH3 binding partner MSH2 (Lopez Castel et al., 2010, Manley et al., 1999), and MutL proteins MLH1 (Pinto et al., 2013a, Lee et al., 2017, Hensman Moss et al., 2017b), PMS2 (Gomes-Pereira, 2004, Gomes-Pereira et al., 2014b, Pinto et al., 2013b) and MLH3 (Pinto et al., 2013a), all of which have been implicated in repeat instability. Each component of this somatic instability network (SIN) will be studied for independent effects on instability and cell phenotype using a panel of shRNA oligonucleotides, as shown in Chapter 5. To assess for gain of function activities, expression vectors encoding proteins modulating repeat stability will be introduced in parallel.

9.2.2 How FAN1 stabilises the CAG repeat

9.2.2.1 Cell models

Building on the work in chapter 6, U20S cells will be complemented with *FAN1* constructs containing the p.C44A/C47A ubiquitin domain inactivating mutation (Jin and Cho, 2017, Thongthip et al., 2016) and the p.L477P DNA binding domain mutant (Smogorzewska et al., 2010a), as well as truncated forms lacking other protein domains (Smogorzewska et al., 2010a). Studies suggest N-terminal constructs, which retain epitopes for the FS2 anti-FAN1 goat antibody, correctly localise to the nucleus and can be used to investigate the function of FAN1 domains (Smogorzewska et al., 2010a). This will allow investigation of which domains, including DNA binding, UBZ and protein interaction, are required for stabilisation of the CAG repeat.

9.2.2.2 Protein interactions

The U20S system allows regulated expression of tagged FAN1, and through co-immunoprecipitation the effect of variation on protein interactions can be probed. The nature of the FAN1-MLH1 interaction will be further studied by chemical cross-linking mass spectrometry (CL-MS), in collaboration with Dr Kostas Thalassinos (UCL).

9.2.2.3 DNA binding

Building on chapter 6, CAG length-dependent binding will be assessed in ChIP fractions by TapeStation and Bioanalyzer (Agilent). ChIP will be also be used to investigate whether FAN1 variants influence its DNA binding, whether other SIN components bind *HTT* CAG repeat DNA and how their occupancy is affected by FAN1 variation. This will answer the question of whether FAN1 regulates access of SIN components to DNA.

9.2.2.4 Mechanism

To investigate the mechanism by which FAN1 regulates instability, cells will be probed using host cell reactivation assays, which involve transfection with reporter plasmids containing a range of DNA structures (Nagel et al., 2014). The role of abnormal DNA structures such as R-loops, which are RNA:DNA hybrids that form during transcription and have been linked with repeat instability (Reddy et al., 2014, Freudenreich, 2018, Su and Freudenreich, 2017), will be assessed using the S9.6 DNA:RNA specific antibody (Merck). The presence of R-loops in the CAG repeat itself can be tested by a modified ChIP procedure termed DRIP (DNA-RNA immunoprecipitation) (Yu et al., 2006), where CAG repeat primers amplify DNA

immunoprecipitated using S9.6. To study the requirement for DNA replication, the cell cycle will be arrested chemically and genetically in unstable cell lines (Gomes-Pereira et al., 2001).

9.2.2.5 p.R507H

The p.R507H FAN1 variant, which is associated with fast progression (GeM-HD, 2015), lies in the SAP DNA binding domain (Jin and Cho, 2017), but preliminary studies in U20S and patient-derived lymphoblasts have found no functional differences in DNA binding or CAG stabilisation. Its impact on protein interactions will be investigated, and iPSCs have been generated from a fast progressing Track-HD patient expressing the variant, which will allow study of its effects in more physiological, neuronally differentiated cells. Structural analysis reveals p.R507H lies near the surface of the SAP domain, a region important for FAN1 dimerisation (Wang et al., 2014b), so its oligomerisation state and protein complexing will be investigated by mass spectrometry.

9.2.3 How N-terminal *MSH3* variation slows disease course

9.2.3.1 Cell models

MSH3 knockout U20S cells will be generated by CRISPR, as described by Munoz et al. (2014), then cells will be complemented with GeneArt synthesised Myc-tagged constructs encoding repeat region variants cloned into the pcDNA5 FRT/TO tetracycline-inducible expression vector, along with pathogenic *HTT* exon 1, as demonstrated in chapter 6. These will be complemented by CRISPR knockout of *MSH3* in 109Q and 125Q iPSCs, currently under way in collaboration with UCL Cancer Institute's Genomics and Genome Engineering department. These iPSCs can differentiate into MSNs and cortical neurons and, as described in chapter 5, demonstrate robust repeat expansion. Sanger sequencing in both lines has shown them to be wild type for the *FAN1* and *MSH3* variants implicated in HD. Whole genome sequencing is under way to fully characterise their genetic background. These lines can then be used to assess the effect of MSH3 level on CAG repeat expansion. Several patient-derived cell lines homozygous for the relatively common 3a repeat allele are also available from the Track-HD and MTM studies.

9.2.3.2 Protein structure

The primary protein structure of the 3a *MSH3* repeat allele will be confirmed by tandem mass spectrometry using protein immunoprecipitated from homozygous cell lines using an MSH3-specific antibody (BD Biosciences).

9.2.3.3 Protein interactions

As the *MSH3* repeat variant lies close to domains interacting with PCNA, EXO1 and MSH2, its effect on protein interactions will be investigated by co-immunoprecipitation using an MSH3 antibody in patient-derived LBs and Myc-trap beads from U20S cells, followed by mass spectrometry to identify and quantify binding partners.

9.2.3.4 Mismatch repair

The effect of repeat variation on MSH3 mismatch repair function will be assessed by sensitivity of patient-derived LB cells to genotoxins including cisplatin (MacKay et al., 2010a, Kratz et al., 2010b), 6-thioguanine (Karran and Attard, 2008, Swann et al., 1996) and hydrogen peroxide (Driessens et al., 2009), and the induced protein interactions will be probed by co-immunoprecipitation, subcellular fractionation and ChIP.

9.2.3.5 DNA interaction

ChIP using highly specific Myc-Trap beads (ChromoTec) will be used to quantify the interaction between MSH3 and *HTT* CAG repeat DNA.

9.2.4 Sequelae of repeat expansion

The Opera high content screening platform will be used to assess how HD cell phenotypes including HTT aggregation using S830 and EM48 antibodies, cell morphology, differentiation, synaptogenesis and mitochondrial function correlate with CAG repeat instability in patient-derived MSNs. To investigate whether genetic variation in SIN components causes widespread DNA damage, oxidative lesions and strand breaks will be studied using the 53BP1, γ-H2AX and the Comet assays (Brierley and Martin, 2013, Niedernhofer et al., 2004, Kurashige et al., 2016).

9.2.5 Summary

This thesis determined that FAN1 and MSH3 form a key part of a DNA repair network that regulates repeat stability, thereby modifying the course of repeat expansion disease. Their modulation has significant therapeutic potential in several of the commonest genetic neurodegenerative diseases (Paulson, 2018). The planned work presented above can be summarised by posing the following questions.

1. How does FAN1 stabilise trinucleotide repeat tracts independent of its nuclease activity?
2. How does N-terminal genetic variation in *MSH3* limit its potentiation of repeat expansion?
3. Along with FAN1 and MSH3, which DNA repair components are involved in the network that regulates repeat stability?
4. Does reducing somatic expansion through therapeutic modulation of this network slow disease course?

The answers to these questions will advance our understanding of repeat expansion disease pathogenesis, and will have important implications for future therapeutic approaches.

Chapter 10 Appendix

10.1 p'HRsincpptUCOE+htt exon1 IRES eGFP 129CAG vector sequence

TCGACCTCGAGGTCGATCCCATGGCGACCCTGGAAAAGCTGATGAAGGCCTTCGAGTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG
CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA
GCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGC
AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG
CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCCGCCACCGCCGCCGCCGCCGCCTCCTCAGCTTCC
TCAGCCGCCGCCGCAGGCACAGCCGCTGCTGCCTCAGCCGCAGCCGCCCCCGCCGCCGCCCCCGCCGCCACCCGGCCCGGCTGTGGCTGAGGAGCCGCTGC
ACCGACCGTGAGTTTGGGCCCGCTGCAGTCGACGGTACCGCGGGCCCGGGATCCGCCCCTCTCCCTCCCCCCCCCCTAACGTTACTGGCCGAAGCCGCTTG
GAATAAGGCCGGTGTGCGTTTGTCTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTGGCCCTGTCTTCTTGACGAG
CATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAGGTCTGTTGAATGTCGTGAAGGAAGCAGTTCCTCTGGAAGCTTCTTGAAGACAAACAACGT
CTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGTGCCTCTGCGGCCAAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAA
CCCCAGTGCCACGTTGTGAGTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGAAGGATGCCCAGAAGGTACCCCA
TTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTACATGTGTTTAGTCGAGGTTAAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTT
TCCTTTGAAAAACACGATGATAATATGGCCACAACCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGA
CGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCG
TGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATG
CCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCG
CATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACA
AGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGC
GACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTT
CGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAAAGCGGCCGCCATATGATAGTTTAAACATTACGGAACTTCGATAATCAACCTC
TGGATTACAAAATTTGTGAAAGATTGACTGGTATTCTTAACTATGTTGCTCCTTTTACGCTATGTGGATACGCTGCTTTAATGCCTTTGTATCATGCTATT
GCTTCCCGTATGGCTTTCATTTTCTCCTCCTTGTATAAATCCTGGTTGCTGTCTCTTTATGAGGAGTTGTGGCCCGTTGTCAGGCAACGTGGCGTGGTGTG
CACTGTGTTTGCTGACGCAACCCCCACTGGTTGGGGCATTGCCACCACCTGTCAGCTCCTTTCCGGGACTTTCGCTTTCCCCCTCCCTATTGCCACGGCGG
AACTCATCGCCGCCTGCCTTGCCCGCTGCTGGACAGGGGCTCGGCTGTTGGGCACTGACAATTCCGTGGTGTTGTCGGGGAAATCATCGTCCTTTCCTTGG
CTGCTCGCCTGTGTTGCCACCTGGATTCTGCGCGGGACGTCCTTCTGCTACGTCCCTTCGGCCCTCAATCCAGCGGACCTTCCTTCCCGCGGCCTGCTGCC
GGCTCTGCGGCCTCTTCCGCGTCTTCGCCTTCGCCCTCAGACGAGTCGGATCTCCCTTTGGGCCGCCTCCCCGCATCGAATCTATGCAGATTCGAGGAATT
AATTCGAGCTCGGTACCTTTAAGACCAATGACTTACAAGGCAGCTGTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCTAATTCACTC
CCAACGAAGACAAGATCTGCTTTTTGCTTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTA
AGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGG
AAAATCTCTAGCAGCATCTAGAATTAATTCCGTGTATTCTATAGTGTCACCTAAATCGTATGTGTATGATACATAAGGTTATGTATTAATTGTAGCCGCGT
TCTAACGACAATATGTACAAGCCTAATTGTGTAGCATCTGGCTTACTGAAGCAGACCCTATCATCTCTCTCGTAAACTGCCGTCAGAGTCGGTTTGGTTGG
ACGAACCTTCTGAGTTTCTGGTAACGCCGTCCCGCACCCGGAAATGGTCAGCGAACCAATCAGCAGGGTCATCGCTAGCCAGATCCTCTACGCCGGACGCA
TCGTGGCCGGCATCACCGGCGCCACAGGTGCGGTTGCTGGCGCCTATATCGCCGACATCACCGATGGGGAAGATCGGGCTCGCCACTTCGGGCTCATGAGC
GCTTGTTTCGGCGTGGGTATGGTGGCAGGCCCCGTGGCCGGGGGACTGTTGGGCGCCATCTCCTTGCATGCACCATTCCTTGCGGCGGCGGTGCTCAACGG
CCTCAACCTACTACTGGGCTGCTTCCTAATGCAGGAGTCGCATAAGGGAGAGCGTCGAATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTA
AGCCAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTG
CATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATGGTTT
CTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAA
CCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGT
TTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCC
TTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAA
CTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAG
TGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATC
ATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGC
AAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCT
TCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCG
TAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGAC
CAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCC
TTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAA
CAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAAT
ACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGC
CAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCA
GCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCATTGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTA
AGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCG
TCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACA
TGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAG
TCAGTGAGCGAGGAAGCGGAAGAGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGTGGAATGTGTGTCAGTTAGGG
TGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGG
CAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTC
CGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTA
GGCTTTTGCAAAAAGCTTGGACACAAGACAGGCTTGCGAGATATGTTTGAGAATACCACTTTATCCCGCGTCAGGGAGAGGCAGTGCGTAAAAAGACGCGG
ACTCATGTGAAATACTGGTTTTTAGTGCGCCAGATCTCTATAATCTCGCGCAACCTATTTTCCCCTCGAACACTTTTAAGCCGTAGATAAACAGGCTGGG
ACACTTCACATGAGCGAAAAATACATCGTCACCTGGGACATGTTGCAGATCCATGCACGTAAACTCGCAAGCCGACTGATGCCTTCTGAACAATGGAAAGG
CATTATTGCCGTAAGCCGTGGCGGTCTGTACCGGGTGCGTTACTGGCGCGTGAACTGGGTATTCGTCATGTCGATACCGTTTGTATTTCCAGCTACGATCA
CGACAACCAGCGCGAGCTTAAAGTGCTGAAACGCGCAGAAGGCGATGGCGAAGGCTTCATCGTTATTGATGACCTGGTGGATACCGGTGGTACTGCGGTTG
CGATTCGTGAAATGTATCCAAAAGCGCACTTTGTCACCATCTTCGCAAAACCGGCTGGTCGTCCGCTGGTTGATGACTATGTTGTTGATATCCCGCAAGAT
ACCTGGATTGAACAGCCGTGGGATATGGGCGTCGTATTCGTCCCGCCAATCTCCGGTCGCTAATCTTTTCAACGCCTGGCACTGCCGGGCGTTGTTCTTTT
TAACTTCAGGCGGGTTACAATAGTTTCCAGTAAGTATTCTGGAGGCTGCATCCATGACACAGGCAAACCTGAGCGAAACCCTGTTCAAACCCCGCTTTAAA
CATCCTGAAACCTCGACGCTAGTCCGCCGCTTTAATCACGGCGCACAACCGCCTGTGCAGTCGGCCCTTGATGGTAAAACCATCCCTCACTGGTATCGCAT
GATTAACCGTCTGATGTGGATCTGGCGCGGCATTGACCCACGCGAAATCCTCGACGTCCAGGCACGTATTGTGATGAGCGATGCCGAACGTACCGACGATG
ATTTATACGATACGGTGATTGGCTACCGTGGCGGCAACTGGATTTATGAGTGGGCCCCGGATCTTTGTGAAGGAACCTTACTTCTGTGGTGTGACATAATT
GGACAAACTACCTACAGAGATTTAAAGCTCTAAGGTAAATATAAAATTTTTAAGTGTATAATGTGTTAAACTACTGATTCTAATTGTTTGTGTATTTTAGA

TTCCAACCTATGGAACTGATGAATGGGAGCAGTGGTGGAATGCCTTTAATGAGGAAAACCTGTTTTGCTCAGAAGAAATGCCATCTAGTGATGATGAGGCT
ACTGCTGACTCTCAACATTCTACTCCTCCAAAAAAGAAGAGAAAGGTAGAAGACCCCAAGGACTTTCCTTCAGAATTGCTAAGTTTTTTGAGTCATGCTGT
GTTTAGTAATAGAACTCTTGCTTGCTTTGCTATTTACACCACAAAGGAAAAAGCTGCACTGCTATACAAGAAAATTATGGAAAAATATTCTGTAACCTTTA
TAAGTAGGCATAACAGTTATAATCATAACATACTGTTTTTTCTTACTCCACACAGGCATAGAGTGTCTGCTATTAATAACTATGCTCAAAAATTGTGTACC
TTTAGCTTTTTAATTTGTAAAGGGGTTAATAAGGAATATTTGATGTATAGTGCCTTGACTAGAGATCATAATCAGCCATACCACATTTGTAGAGGTTTTAC
TTGCTTTAAAAAACCTCCCACACCTCCCCCTGAACCTGAAACATAAAATGAATGCAATTGTTGTTGTTAACTTGTTTATTGCAGCTTATAATGGTTACAAA
TAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGGAT
CAACTGGATAACTCAAGCTAACCAAAATCATCCCAAACTTCCCACCCCATACCCTATTACCACTGCCAATTACCTAGTGGTTTCATTTACTCTAAACCTGT
GATTCCTCTGAATTATTTTCATTTTAAAGAAATTGTATTTGTTAAATATGTACTACAAACTTAGTAGTTGGAAGGGCTAATTCACTCCCAAAGAAGACAAG
ATATCCTTGATCTGTGGATCTACCACACACAAGGCTACTTCCCTGATTAGCAGAACTACACACCAGGGCCAGGGGTCAGATATCCACTGACCTTTGGATGG
TGCTACAAGCTAGTACCAGTTGAGCCAGATAAGGTAGAAGAGGCCAATAAAGGAGAGAACACCAGCTTGTTACACCCTGTGAGCCTGCATGGGATGGATGA
CCCGGAGAGAGAAGTGTTAGAGTGGAGGTTTGACAGCCGCCTAGCATTTCATCACGTGGCCCGAGAGCTGCATCCGGAGTACTTCAAGAACTGCTGATATC
GAGCTTGCTACAAGGGACTTTCCGCTGGGGACTTTCCAGGGAGGCGTGGCCTGGGCGGGACTGGGGAGTGGCGAGCCCTCAGATCCTGCATATAAGCAGCT
GCTTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCCACTGCTTAAGCCTCAATAAAGCTTGC
CTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGTGTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTAGCAGTGGC
GCCCGAACAGGGACTTGAAAGCGAAAGGGAAACCAGAGGAGCTCTCTCGACGCAGGACTCGGCTTGCTGAAGCGCGCACGGCAAGAGGCGAGGGGCGGCGA
CTGGTGAGTACGCCAAAAATTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAGTATTAAGCGGGGGAGAATTAGATCGCGATGGGA
AAAAATTCGGTTAAGGCCAGGGGGAAAGAAAAAATATAAATTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAACGATTCGCAGTTAATCCTGGCCTGT
TAGAAACATCAGAAGGCTGTAGACAAATACTGGGACAGCTACAACCATCCCTTCAGACAGGATCAGAAGAACTTAGATCATTATATAATACAGTAGCAACC
CTCTATTGTGTGCATCAAAGGATAGAGATAAAAGACACCAAGGAAGCTTTAGACAAGATAGAGGAAGAGCAAAACAAAAGTAAGACCACCGCACAGCAAGC
GGCCGCTGATCTTCAGACCTGGAGGAGGAGATATGAGGGACAATTGGAGAAGTGAATTATATAAATATAAAGTAGTAAAAATTGAACCATTAGGAGTAGCA
CCCACCAAGGCAAAGAGAAGAGTGGTGCAGAGAGAAAAAAGAGCAGTGGGAATAGGAGCTTTGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACTATGGG
CGCAGCGTCAATGACGCTGACGGTACAGGCCAGACAATTATTGTCTGGTATAGTGCAGCAGCAGAACAATTTGCTGAGGGCTATTGAGGCGCAACAGCATC
TGTTGCAACTCACAGTCTGGGGCATCAAGCAGCTCCAGGCAAGAATCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTCCTGGGGATTTGGGGTTGC
TCTGGAAAACTCATTTGCACCACTGCTGTGCCTTGGAATGCTAGTTGGAGTAATAAATCTCTGGAACAGATTTGGAATCACACGACCTGGATGGAGTGGGA
CAGAGAAATTAACAATTACACAAGCTTAATACACTCCTTAATTGAAGAATCGCAAAACCAGCAAGAAAAGAATGAACAAGAATTATTGGAATTAGATAAAT
GGGCAAGTTTGTGGAATTGGTTTAACATAACAAATTGGCTGTGGTATATAAAATTATTCATAATGATAGTAGGAGGCTTGGTAGGTTTAAGAATAGTTTTT
GCTGTACTTTCTATAGTGAATAGAGTTAGGCAGGGATATTCACCATTATCGTTTCAGACCCACCTCCCAACCCCGAGGGGACCCGACAGGCCCGAAGGAAT
AGAAGAAGAAGGTGGAGAGAGAGACAGAGACAGATCCATTCGATTAGTGAACGGATCTCGACGGTATCGCCAAATGGCAGTATTCATCCACAATTTTAAAA
GAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAACAAATTACAAAAATT
CAAAATTTTCGGGTTTATTACAGGGACAGCAGAGATCCAGTTTGGATTGATAAGCTTGATATCGAATTCGTTTTTCTCTTAATTTTCCCCTGTAATTTACA
CTGGGAGAGCTGGGAAATATGTGGATGTAAATTTCTCAGCCACAGAGATGCAAAGTTATACTGTGGGGAAAAAAAAACTTGAGTTAAATCCTTACATATTTT
AGGTTTTCATTAACTTACCAATGTAGTTTTGTTGGAGGCCATTTTTTATTGCAGACTTGAAGAGCTATTACTAGAAAAATGCATGACAGTTAAGGTAAGTT
TGCATGACACAAAAAAGGTAACTAAATACAAATTCTGTTTGGATTCCAACCCCCAAGTAGAGAGCGCACACTTTCAAACGTGAATACAAATCCAGAGTAGA
TCTGCGCTCCTACCTACATTGCTTATGATGTACTTAAGTACGTGTCCTAACCATGTGAGTCTAGAAAGACTTTACTGGGGATCCTGGTACCTAAAACAGCT
TCACATGGCTTAAAATAGGGGACCAATGTCTTTTCCAATCTAAGTCCCATTTATAATAAAGTCCATGTTCCATTTTTAAAGGACAATCCTTTCGGTTTAAA
ACCAGGCACGATTACCCAAACAACTCACAACGGTAAAGCACTGTGAATCTTCTCTGTTCTGCAATCCCAACTTGGTTTCTGCTCAGAAACCCTCCCTCTTT
CCAATCGGTAATTAAATAACAAAAGGAAAAAACTTAAGATGCTTCAACCCCGTTTCGTGACACTTTGAAAAAAGAATCACCTCTTGCAAACACCCGCTCCC
GACCCCCGCCGCTGAAGCCCGGCGTCCAGAGGCCTAAGCGCGGGTGCCCGCCCCCACCCGGGAGCGCGGGCCTCGTGGTCAGCGCATCCGCGGGGAGAAAC
AAAGGCCGCGGCACGGGGGCTCAAGGGCACTGCGCCACACCGCACGCGCCTACCCCCGCGCGGCCACGTTAACTGGCGGTCGCCGCAGCCTCGGGACAGCC
GGCCGCGCGCCGCCAGGCTCGCGGACGCGGGACCACGCGCCGCCCTCCGGGAGGCCCAAGTCTCGACCCAGCCCCGCGTGGCGCTGGGGGAGGGGGCGCCT
CCGCCGGAACGCGGGTGGGGGAGGGGAGGGGGAAATGCGCTTTGTCTCGAAATGGGGCAACCGTCGCCACAGCTCCCTACCCCCTCGAGGGCAGAGCAGTC
CCCCCACTAACTACCGGGCTGGCCGCGCGCCAGGCCAGCCGCGAGGCCACCGCCCGACCCTCCACTCCTTCCCGCAGCTCCCGGCGCGGGGTCCGGCGAGA
AGGGGAGGGGAGGGGAGCGGAGAACCGGGCCCCCGGGACGCGTGTGGCATCTGAAGCACCACCAGCGAGCGAGAGCTAGAGAGAAGGAAAGCCACCGACTT
CACCGCCTCCGAGCTGCTCCGGGTCGCGGGTCTGCAGCGTCTCCGGCCCTCCGCGCCTACAGCTCAAGCCACATCCGAAGGGGGAGGGAGCCGGGAGCTGC
GCGCGGGGCCGCCGGGGGGAGGGGTGGCACCGCCCACGCCGGGCGGCCACGAAGGGCGGGGCAGCGGGCGCGCGCGGCGGGGGGAGGGGCCGGCGCCGC
GCCCGCTGGGAATTGGGGCCCTAGGGGGAGGGCGGAGGCGCCGACGACCGCGGCACTTACCGTTCGCGGCGTGGCGCCCGGTGGTCCCCAAGGGGAGGGAA
GGGGGAGGCGGGGCGAGGACAGTGACCGGAGTCTCCTCAGCGGTGGCTTTTCTGCTTGGCAGCCTCAGCGGCTGGCGCCAAAACCGGACTCCGCCCACTTC
CTCGCCCGCCGGTGCGAGGGTGTGGAATCCTCCAGACGCTGGGGGAGGGGGAGTTGGGAGCTTAAAAACTAGTACCCCTTTGGGACCACTTTCAGCAGCGA
ACTCTCCTGTACACCAGGGGTCAGTTCCACAGACGCGGGCCAGGGGTGGGTCATTGCGGCGTGAACAATAATTTGACTAGAAGTTGATTCGGGTGTTTCCG
GAAGGGGCCGAGTCAATCCGCCGAGTTGGGGCACGGAAAACAAAAAGGGAAGGCTACTAAGATTTTTCTGGCGGGGGTTATCATTGGCGTAACTGCAGGGA
CCACCTCCCGGGTTGAGGGGGCTGGATCTCCAGGCTGCGGATTAAGCCCCTCCCGTCGGCGTTAATTTCAAACTGCGCGACGTTTCTCACCTGCCTTCGCC
AAGGCAGGGGCCGGGACCCTATTCCAAGAGGTAGTAACTAGCAGGACTCTAGCCTTCCGCAATTCATTGAGCGCATTTACGGAAGTAACGTCGGGTACTGT
CTCTGGCCGCAAGGGTGGGAGGAGTACGCATTTGGCGTAAGGTGGGGCGTAGAGCCTTCCCGCCATTGGCGGCGGATAGGGCGTTTACGCGACGGCCTGAC
GTAGCGGAAGACGCGTTAGTGGGGGGGAAGGTTCTAGAAAAGCGGCGGCAGCGGCTCTAGCGGCAGTAGCAGCAGCGCCGGGTCCCGTGCGGAGGTGCTCC
TCGCAGAGTTGTTTCTCGAGCAGCGGCAGTTCTCACTACAGCGCCAGGACGAGTCCGGTTCGTGTTCGTCCGCGGAGATCTCTCTCATCTCGCTCGGCTGC
GGGAAATCGGGCTGAAGCGACTGAGTCCCCGGGTCTAGAATCGATAAGCTTGAGCTCGATATCG

10.2 U20S curve modelling in R

```
# Get data

series <- read_excel('data.xlsx')

# Change the shape of the data to long format

series.long <- series %>%

  gather(Q, value, 2:5) %>%

  mutate(Q = factor(Q, levels=c('30Q', '70Q', '97Q', '118Q')))

na.omit() %>%

  as.data.frame()

# Fit the model (you can change the model here)

expmod_118Q <- lm(log(`118Q`) ~ day, data=series)

expmod_97Q <- lm(log(`97Q`) ~ day, data=series)

# Make predictions

prediction.data_118Q <- data.frame(day=1:3000)

predictions_118Q <- predict(expmod_118Q, prediction.data_118Q, interval = "c", level=0.95) %>% as.data.frame() %>% mutate(day=1:3000)

prediction.data_97Q <- data.frame(day=1:3000)

predictions_97Q <- predict(expmod_97Q, prediction.data_97Q, interval = "c", level=0.95) %>% as.data.frame() %>% mutate(day=1:3000)

# Change the fit to normal value not log

predictions_118Q[, 1:3] <- exp(predictions_118Q[, 1:3])

predictions_97Q[, 1:3] <- exp(predictions_97Q[, 1:3])

# Plot values and predictions

ggplot(series.long, aes(day, value)) +

  geom_smooth(aes(y=fit, ymin=lwr, ymax=upr), data=predictions_118Q, stat="identity", color="Purple", fill="Purple") +

  geom_smooth(aes(y=fit, ymin=lwr, ymax=upr), data=predictions_97Q, stat="identity", color="Turquoise", fill="Turquoise") +

  geom_point(aes(col=Q)) +

  theme_bw() +

  labs(x="Day", y="CAG") +

  theme(text = element_text(size=20))

# Filter the data

series.long %>%

  filter(Q %in% c('97Q', '118Q')) %>%

  ggplot(aes(day, value)) +

  geom_smooth(aes(y=fit, ymin=lwr, ymax=upr), data=predictions_118Q, stat="identity", color="Purple", fill="Purple") +

  geom_smooth(aes(y=fit, ymin=lwr, ymax=upr), data=predictions_97Q, stat="identity", color="Turquoise", fill="Turquoise") +

  geom_point(aes(col=Q)) +

  xlim(1700, 2500) +

  theme_bw()
```

10.3 pSUPER.retro.puro vector sequence

TGAAAGACCCCACCTGTAGGTTTGGCAAGCTAGCTTAAGTAACGCCATTTTGCAAGGCATGGAAAATACATAACTGAGAATAGAGAAGTTCAGATCAAGGT
TAGGAACAGAGAGACAGCAGAATATGGGCCAAACAGGATATCTGTGGTAAGCAGTTCCTGCCCCGGCTCAGGGCCAAGAACAGATGGTCCCCAGATGCGGT
CCCGCCCTCAGCAGTTTCTAGAGAACCATCAGATGTTTCCAGGGTGCCCCAAGGACCTGAAATGACCCTGTGCCTTATTTGAACTAACCAATCAGTTCGCT
TCTCGCTTCTGTTCGCGCGCTTCTGCTCCCCGAGCTCAATAAAAGAGCCCACAACCCCTCACTCGGCGCGCCAGTCCTCCGATAGACTGCGTCGCCCGGGT
ACCCGTATTCCCAATAAAGCCTCTTGCTGTTTGCATCCGAATCGTGGACTCGCTGATCCTTGGGAGGGTCTCCTCAGATTGATTGACTGCCCACCTCGGGG
GTCTTTCATTTGGAGGTTCCACCGAGATTTGGAGACCCCTGCCCAGGGACCACCGACCCCCCCGCCGGGAGGTAAGCTGGCCAGCGGTCGTTTCGTGTCTG
TCTCTGTCTTTGTGCGTGTTTGTGCCGGCATCTAATGTTTGCGCCTGCGTCTGTACTAGTTAGCTAACTAGCTCTGTATCTGGCGGACCCGTGGTGGAACT
GACGAGTTCTGAACACCCGGCCGCAACCCTGGGAGACGTCCCAGGGACTTTGGGGGCCGTTTTTGTGGCCCGACCTGAGGAAGGGAGTCGATGTGGAATCC
GACCCCGTCAGGATATGTGGTTCTGGTAGGAGACGAGAACCTAAAACAGTTCCCGCCTCCGTCTGAATTTTTGCTTTCGGTTTGGAACCGAAGCCGCGCGT
CTTGTCTGCTGCAGCGCTGCAGCATCGTTCTGTGTTGTCTCTGTCTGACTGTGTTTCTGTATTTGTCTGAAAATTAGGGCCAGACTGTTACCACTCCCTTA
AGTTTGACCTTAGGTCACTGGAAAGATGTCGAGCGGATCGCTCACAACCAGTCGGTAGATGTCAAGAAGAGACGTTGGGTTACCTTCTGCTCTGCAGAATG
GCCAACCTTTAACGTCGGATGGCCGCGAGACGGCACCTTTAACCGAGACCTCATCACCCAGGTTAAGATCAAGGTCTTTTCACCTGGCCCGCATGGACACC
CAGACCAGGTCCCCTACATCGTGACCTGGGAAGCCTTGGCTTTTGACCCCCCTCCCTGGGTCAAGCCCTTTGTACACCCTAAGCCTCCGCCTCCTCTTCCT
CCATCCGCCCCGTCTCTCCCCCTTGAACCTCCTCGTTCGACCCCGCCTCGATCCTCCCTTTATCCAGCCCTCACTCCTTCTCTAGGCGCCGGAATTAGATC
GATCTCTCGAGGTCGACGGTATCGATAAGCTTAGATCTGTGGTCTCATACAGAACTTATAAGATTCCCAAATCCAAAGACATTTCACGTTTATGGTGATTT
CCCAGAACACATAGCGACATGCAAATATTGCAGGGCGCCACTCCCCTGTCCCTCACAGCCATCTTCCTGCCAGGGCGCACGCGCGCTGGGTGTTCCCGCCT
AGTGACACTGGGCCCGCGATTCCTTGGAGCGGGTTGATGACGTCAGCGTTCGAATTCTACCGGGTAGGGGAGGCGCTTTTCCCAAGGCAGTCTGGAGCATG
CGCTTTAGCAGCCCCGCTGGGCACTTGGCGCTACACAAGTGGCCTCTGGCCTCGCACACATTCCACATCCACCGGTAGGCGCCAACCGGCTCCGTTCTTTG
GTGGCCCCTTCGCGCCACCTTCTACTCCTCCCCTAGTCAGGAAGTTCCCCCCCGCCCCGCAGCTCGCGTCGTGCAGGACGTGACAAATGGAAGTAGCACGT
CTCACTAGTCTCGTGCAGATGGACAGCACCGCTGAGCAATGGAAGCGGGTAGGCCTTTGGGGCAGCGGCCAATAGCAGCTTTGCTCCTTCGCTTTCTGGGC
TCAGAGGCTGGGAAGGGGTGGGTCCGGGGGCGGGCTCAGGGGCGGGCTCAGGGGCGGGGCGGGCGCCCGAAGGTCCTCCGGAGGCCCGGCATTCTGCACGC
TTCAAAAGCGCACGTCTGCCGCGCTGTTCTCCTCTTCCTCATCTCCGGGCCTTTCGACCTGCAGCCCAAGCTAGCTTACCATGACCGAGTACAAGCCCACG
GTGCGCCTCGCCACCCGCGACGACGTCCCCAGGGCCGTACGCACCCTCGCCGCCGCGTTCGCCGACTACCCCGCCACGCGCCACACCGTCGATCCGGACCG
CCACATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCGCGTCGGGCTCGACATCGGCAAGGTGTGGGTCGCGGACGACGGCGCCGCGGTGGCGG
TCTGGACCACGCCGGAGAGCGTCGAAGCGGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTTGAGCGGTTCCCGGCTGGCCGCGCAGCAACAG
ATGGAAGGCCTCCTGGCGCCGCACCGGCCCAAGGAGCCCGCGTGGTTCCTGGCCACCGTCGGCGTCTCGCCCGACCACCAGGGCAAGGGTCTGGGCAGCGC
CGTCGTGCTCCCCGGAGTGGAGGCGGCCGAGCGCGCCGGGGTGCCCGCCTTCCTGGAGACCTCCGCGCCCCGCAACCTCCCCTTCTACGAGCGGCTCGGCT
TCACCGTCACCGCCGACGTCGAGGTGCCCGAAGGACCGCGCACCTGGTGCATGACCCGCAAGCCCGGTGCCTGACGCCCGCCCCACGACCCGCAGCGCCCG
ACCGAAAGGAGCGCACGACCCCATGCATCGATAAAATAAAAGATTTTATTTAGTCTCCAGAAAAAGGGGGGAATGAAAGACCCCACCTGTAGGTTTGGCAA
GCTAGAGAACCATCAGATGTTTCCAGGGTGCCCCAAGGACCTGAAATGACCCTGTGCCTTATTTGAACTAACCAATCAGTTCGCTTCTCGCTTCTGTTCGC
GCGCTTCTGCTCCCCGAGCTCAATAAAAGAGCCCACAACCCCTCACTCGGCGCGCCAGTCCTCCGATAGACTGCGTCGCCCGGGTACCCGTGTATCCAATA
AACCCTCTTGCAGTTGCATCCGACTTGTGGTCTCGCTGTTCCTTGGGAGGGTCTCCTCTGAGTGATTGACTACCCGTCAGCGGGGGTCTTTCATGGGTAAC
AGTTTCTTGAAGTTGGAGAACAACATTCTGAGGGTAGGAGTCGAATATTAAGTAATCCTGACTCAATTAGCCACTGTTTTGAATCCACATACTCCAATACT
CCTGAAATAGTTCATTATGGACAGCGCAGAAGAGCTGGGGAGAATTAATTCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAA
TTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTC
CAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGA
CTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACA
TGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGA
CGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCT
TACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGC
TGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTG
GCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGT
ATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTG
TTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAA
GGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTG
GTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTA
CGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGA
AGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTT
GCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACAT
GATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCA
CTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAG
TTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAA
GGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAA
ACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGG
TTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAG
AAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATG
CAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGGCT
TAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCCATT
CGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGT
TGGGTAACGCCAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGG
CCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGC
ACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGCGATTAGTCCAATTTGTTAAAGACAGGATATCAGTGGTCCAGGCTCTAGTTTTGA
CTCAACAATATCACCAGCTGAAGCCTATAGAGTACGAGCCATAGATAAAATAAAAGATTTTATTTAGTCTCCAGAAAAAGGGGGGAA

10.4 pcDNA5-GFP-FAN1/FRT/TO vector sequence

ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGG
CGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACG
GCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTGCAGGAGCGCACCATCTTCTTC
AAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGG
CAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCC
GCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTG
AGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGA
GCTGTACAAGTCCGGACTCGGATCTATGATGTCAGAAGGGAAACCTCCTGACAAAAAAAGGCCTCGTAGAAGCTTATCAATCAGCAAGAATAAGAAAAAAG
CATCTAATTCTATTATTTCGTGTTTTAACAATGCACCACCTGCTAAACTTGCCTGCCCCGTTTGCAGTAAAATGGTGCCTAGATATGACTTAAACCGGCAC
CTTGATGAAATGTGTGCTAACAATGACTTCGTTCAAGTGGATCCAGGGCAGGTTGGCTTAATAAATTCAAATGTGTCTATGGTAGATTTAACCAGTGTTAC
CTTAGAAGATGTAACACCTAAGAAGTCACCACCACCAAAGACAAATTTAACCCCTGGCCAAAGTGATTCAGCAAAAAGGGAAGTAAAGCAGAAGATCAGTC
CCTACTTTAAAAGTAATGATGTGGTGTGCAAAAATCAAGATGAGCTGAGAAATCGTAGTGTGAAAGTCATTTGTTTGGGAAGCCTAGCATCTAAATTGTCC
AGAAAATACGTAAAGGCTAAAAAATCAATAGATAAGGATGAAGAATTTGCCGGTTCTAGTCCACAGAGTTCCAAATCCACAGTTGTTAAGAGCCTGATTGA
TAACTCTTCAGAAATTGAGGACGAGGATCAAATTTTGGAGAACAGTTCTCAAAAAGAAAACGTGTTTAAATGTGATTCTCTAAAGGAAGAGTGCATTCCTG
AACATATGGTAAGAGGAAGTAAAATAATGGAAGCCGAAAGCCAAAAGGCTACCCGGGAATGTGAGAAATCAGCCCTCACCCCTGGATTCTCAGATAATGCG
ATCATGTTATTCTCACCAGATTTCACTCTTAGGAATACATTAAAGTCTACTTCAGAAGACAGTCTTGTAAAGCAAGAGTGTATCAAAGAAGTGGTTGAAAA
ACGTGAGGCATGTCATTGTGAAGAAGTAAAAATGACTGTTGCTTCAGAAGCTAAAATACAGCTGTCAGATTCAGAGGCAAAATCTCATAGTTCTGCAGATG
ATGCTTCTGCATGGAGTAACATCCAAGAGGCTCCTCTGCAGGATGACAGTTGCTTAAACAATGATATCCCTCACAGCATTCCTTTGGAGCAGGGGTCAAGC
TGCAATGGTCCTGGTCAAACAACCGGTCATCCTTACTACCTTCGGAGTTTCCTTGTGGTGCTGAAAACCGTACTTGAGAATGAAGATGATATGTTGCTCTT
TGATGAGCAGGAGAAGGGAATTGTAACTAAATTTTATCAGTTATCAGCTACTGGTCAGAAGTTATATGTAAGGCTCTTTCAACGTAAATTAAGCTGGATTA
AGATGACCAAATTAGAGTATGAAGAGATTGCCTTAGACTTAACACCTGTGATTGAAGAATTGACGAATGCAGGCTTTCTACAGACAGAATCTGAGTTGCAA
GAACTCTCTGAAGTGCTTGAACTCCTTTCTGCTCCTGAACTAAAATCCCTAGCCAAGACCTTCCACTTGGTGAATCCCAATGGACAGAAACAGCAGCTGGT
GGACGCCTTTCTCAAATTGGCCAAACAGCGTTCAGTCTGCACTTGGGGCAAGAATAAGCCTGGAATTGGTGCAGTGATTTTAAAAAGAGCCAAAGCCTTGG
CTGGACAGTCAGTACGAATCTGTAAAGGCCCCAGGGCTGTGTTTTCCCGCATCTTGCTACTGTTTTCGTTGACCGACTCAATGGAAGATGAAGACGCCGCT
TGTGGAGGTCAGGGACAGCTTTCAACAGTCCTGTTGGTCAACCTCGGCCGAATGGAGTTTCCTAGTTACACCATCAATCGGAAAACCCACATCTTCCAAGA
CAGAGATGATCTTATCAGATATGCAGCAGCCACGCACATGCTGAGTGACATTTCTTCCGCAATGGCCAATGGGAACTGGGAAGAAGCTAAGGAGCTCGCTC
AGTGTGCAAAAAGGGATTGGAACAGACTGAAAAACCACCCTTCTCTGAGATGCCACGAAGATTTACCACTCTTCCTGCGGTGTTTCACTGTTGGGTGGATT
TATACAAGGATTTTGTCTCGGTTTGTGGAAATACTGCAGAGACTTCACATGTATGAGGAAGCCGTCAGAGAACTTGAAAGCCTTTTGTCTCAGAGAATTTA
TTGTCCTGACAGCAGAGGCCGATGGTGGGATCGACTGGCCCTTAATTTACACCAGCACTTGAAGCGCCTGGAACCGACTATCAAGTGCATCACAGAGGGGC
TGGCGGATCCGGAAGTCAGAACGGGACACCGCCTTTCACTGTATCAGCGAGCCGTGCGCCTGCGAGAGTCTCCGAGCTGTAAAAAGTTCAAGCACCTCTTC
CAGCAGCTCCCAGAAATGGCTGTGCAAGATGTGAAACACGTGACCATCACAGGCAGGCTGTGCCCACAGCGTGGGATGTGCAAGTCTGTGTTTGTGATGGA
GGCCGGGGAGGCCGCTGACCCCACCACGGTCCTGTGCTCTGTGGAGGAGCTGGCACTGGCCCATTACAGACGCAGCGGTTTTGACCAGGGGATTCATGGCG
AAGGGTCCACCTTCAGCACCCTGTATGGCCTCCTCCTGTGGGACATCATCTTCATGGATGGGATTCCGGATGTCTTCAGAAACGCCTGTCAGGCATTCCCC
CTGGACTTGTGCACAGACAGCTTCTTCACAAGCAGACGCCCAGCCCTTGAGGCCAGGCTGCAGCTGATTCATGATGCCCCCGAGGAGAGCCTGCGGGCCTG
GGTGGCAGCCACGTGGCATGAGCAGGAAGGCAGAGTGGCTTCCCTTGTCAGCTGGGATCGCTTCACGTCTCTTCAGCAAGCTCAGGATCTTGTCTCCTGCC
TGGGGGGCCCTGTGCTCAGTGGTGTGTGCAGGCACCTGGCTGCTGACTTTCGACACTGTCGAGGGGGCCTCCCCGACCTGGTGGTGTGGAACTCCCAGAGC
CGTCACTTTAAGCTGGTGGAAGTTAAAGGCCCCAATGATCGTCTTTCACATAAGCAGATGATCTGGCTGGCTGAACTGCAGAAGCTGGGGGCTGAAGTAGA
AGTCTGCCACGTGGTTGCAGTTGGAGCTAAG

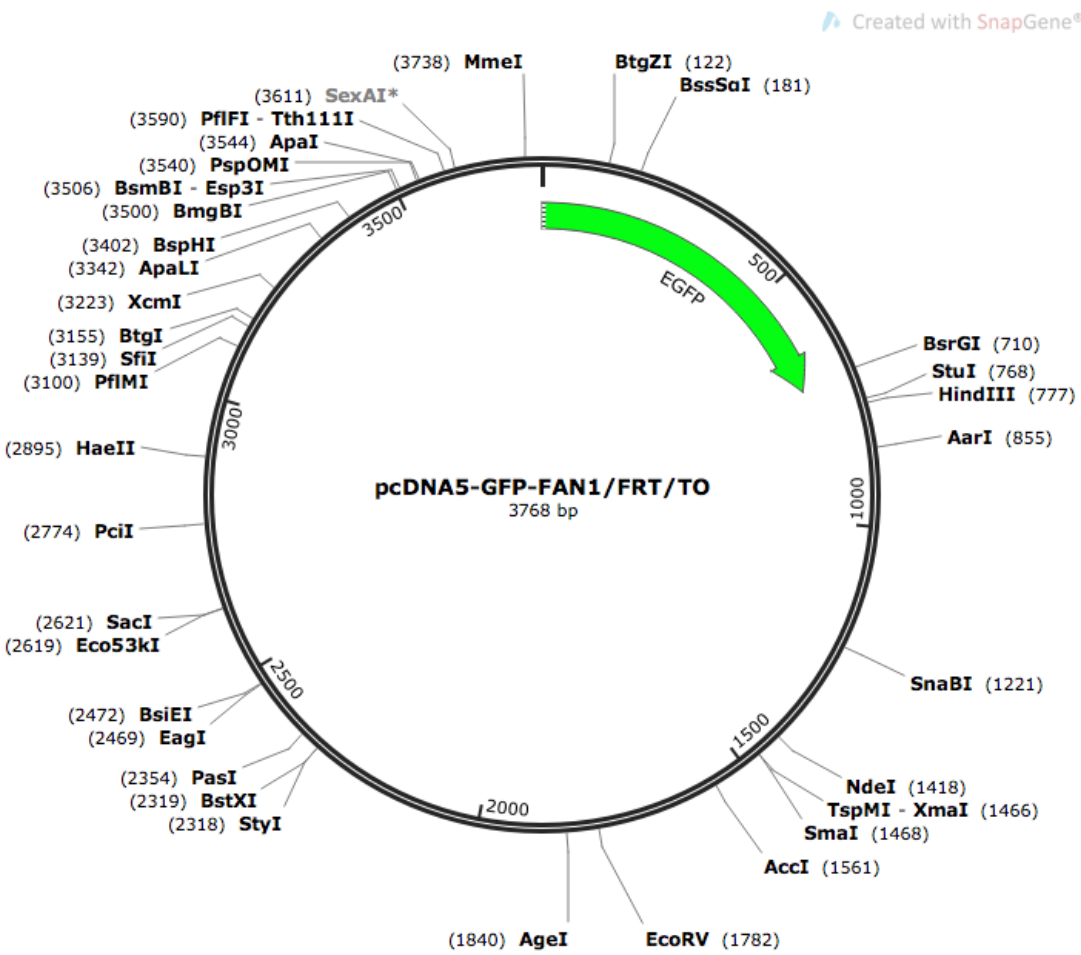


Figure 10.1. Schematic representation of pcDNA5-GFP-FAN1/FRT/TO vector

10.5 Sequencing of A2UCOE *HTT* exon 1 construct CAG repeat regions

Figure 10.2. Sanger sequencing of A2UCOE HTT exon 1 construct CAG repeat regions.
Sequencing conducted by Trager et al. (2014). Sequence aligned to the reference genome (Homo sapiens chromosome 4, GRCh38.p12 Primary Assembly, NCBI Reference Sequence: NC_000004.12) in SnapGene. The polyglutamine repeat is given in red.

10.6 *MSH3* MiSeq primer sequences

	MiSeq primer name	P5 adaptor sequence	i5 index	Sequencing primer binding site	Spacer	Gene-specific primer	Full MiSeq primer sequence
P5 primers (5'-3')	MSH3-F-S502	AATGATACGGCGACCACCGAGATCTACAC	CTCTCTAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT		CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACTCTCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCCTCCTCCAGCCCTATC
	MSH3-F-S503	AATGATACGGCGACCACCGAGATCTACAC	TATCCTCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGA	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACTTCCTCCTCCAGCCCTATC
	MSH3-F-S505	AATGATACGGCGACCACCGAGATCTACAC	GTAAGGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GT	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCTTCCTCCTCCAGCCCTATC
	MSH3-F-S506	AATGATACGGCGACCACCGAGATCTACAC	ACTGCATA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAAG	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACACTGCATAACACTCTTTCCCTACACGACGCTCTTCCGATCTCAAGCTTCCTCCTCCAGCCCTATC
	MSH3-F-S507	AATGATACGGCGACCACCGAGATCTACAC	AAGGAGTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	G	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACAAGGAGTAACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTTCCTCCTCCAGCCCTATC
	MSH3-F-S508	AATGATACGGCGACCACCGAGATCTACAC	CTAAGCCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TCGGA	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGGACTTCCTCCTCCAGCCCTATC
	MSH3-F-S510	AATGATACGGCGACCACCGAGATCTACAC	CGTCTAAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGGAGG	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACCGTCTAATACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGAGGCTTCCTCCTCCAGCCCTATC
	MSH3-F-S511	AATGATACGGCGACCACCGAGATCTACAC	TCTCTCCG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AACAGAC	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACTCTCTCCGACACTCTTTCCCTACACGACGCTCTTCCGATCTAACAGACCTTCCTCCTCCAGCCCTATC
	MSH3-F-S513	AATGATACGGCGACCACCGAGATCTACAC	TCGACTAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT		CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACTCGACTAGACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTCCTCCTCCAGCCCTATC
	MSH3-F-S515	AATGATACGGCGACCACCGAGATCTACAC	TTCTAGCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ATA	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACTTCTAGCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATACTTCCTCCTCCAGCCCTATC
	MSH3-F-S516	AATGATACGGCGACCACCGAGATCTACAC	CCTAGAGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TG	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACCCTAGAGTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCTTCCTCCTCCAGCCCTATC
	MSH3-F-S517	AATGATACGGCGACCACCGAGATCTACAC	GCGTAAGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCAG	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACGCGTAAGAACACTCTTTCCCTACACGACGCTCTTCCGATCTGCAGCTTCCTCCTCCAGCCCTATC
	MSH3-F-S518	AATGATACGGCGACCACCGAGATCTACAC	CTATTAAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	G	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACCTATTAAGACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTTCCTCCTCCAGCCCTATC
	MSH3-F-S520	AATGATACGGCGACCACCGAGATCTACAC	AAGGCTAT	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGCGA	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACAAGGCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCGACTTCCTCCTCCAGCCCTATC
	MSH3-F-S521	AATGATACGGCGACCACCGAGATCTACAC	GAGCCTTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAGAGG	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACGAGCCTTAACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGAGGCTTCCTCCTCCAGCCCTATC
	MSH3-F-S522	AATGATACGGCGACCACCGAGATCTACAC	TTATGCGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAGAGAG	CTTCCTCCTCCAGCCCTATC	AATGATACGGCGACCACCGAGATCTACACTTATGCGAACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGAGAGCTTCCTCCTCCAGCCCTATC

	MiSeq primer name	P7 adaptor sequence	i7 index	Sequencing primer binding site	Spacer	Gene-specific primer	Full MiSeq primer sequence
P5 primers (5'-3')	MSH3-R-N701	CAAGCAGAAGACGGCATACGAGAT	TCGCCTTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTTGGCGCGAAATTGTGG
	MSH3-R-N702	CAAGCAGAAGACGGCATACGAGAT	TAGTACG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTAGTACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGTTTGGCGCGAAATTGTGG
	MSH3-R-N703	CAAGCAGAAGACGGCATACGAGAT	TTCTGCCT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTTGGCGCGAAATTGTGG
	MSH3-R-N704	CAAGCAGAAGACGGCATACGAGAT	GCTCAGGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGTTTGGCGCGAAATTGTGG
	MSH3-R-N705	CAAGCAGAAGACGGCATACGAGAT	AGGAGTCC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGTTTGGCGCGAAATTGTGG
	MSH3-R-N706	CAAGCAGAAGACGGCATACGAGAT	CATGCCTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGTTTGGCGCGAAATTGTGG
	MSH3-R-N707	CAAGCAGAAGACGGCATACGAGAT	GTAGAGAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTTGGCGCGAAATTGTGG
	MSH3-R-N710	CAAGCAGAAGACGGCATACGAGAT	CAGCCTCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGTTTGGCGCGAAATTGTGG
	MSH3-R-N711	CAAGCAGAAGACGGCATACGAGAT	TGCCTCTT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGTTTGGCGCGAAATTGTGG
	MSH3-R-N712	CAAGCAGAAGACGGCATACGAGAT	TCCTCTAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTTGGCGCGAAATTGTGG
	MSH3-R-N714	CAAGCAGAAGACGGCATACGAGAT	TCATGAGC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTCATGAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGTTTGGCGCGAAATTGTGG
	MSH3-R-N715	CAAGCAGAAGACGGCATACGAGAT	CTGAGAT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATCTGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTTGGCGCGAAATTGTGG
	MSH3-R-N716	CAAGCAGAAGACGGCATACGAGAT	TAGCGAGT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTAGCGAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGTTTGGCGCGAAATTGTGG
	MSH3-R-N718	CAAGCAGAAGACGGCATACGAGAT	GTAGCTCC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATGTAGCTCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGTTTGGCGCGAAATTGTGG
	MSH3-R-N719	CAAGCAGAAGACGGCATACGAGAT	TACTACGC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTACTACGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTTGGCGCGAAATTGTGG
	MSH3-R-N720	CAAGCAGAAGACGGCATACGAGAT	GGCTCCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC		AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATGGCTCCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCAGTTTGGCGCGAAATTGTGG
	MSH3-R-N721	CAAGCAGAAGACGGCATACGAGAT	CAGCGTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATCAGCGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGTTTGGCGCGAAATTGTGG
	MSH3-R-N722	CAAGCAGAAGACGGCATACGAGAT	CTGCGCAT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATCTGCGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGTTTGGCGCGAAATTGTGG
	MSH3-R-N723	CAAGCAGAAGACGGCATACGAGAT	GAGCGCTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATGAGCGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGTTTGGCGCGAAATTGTGG
	MSH3-R-N724	CAAGCAGAAGACGGCATACGAGAT	CGCTCAGT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATCGCTCAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGTTTGGCGCGAAATTGTGG
	MSH3-R-N726	CAAGCAGAAGACGGCATACGAGAT	GTCTTAGG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATGTCTTAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGTTTGGCGCGAAATTGTGG
	MSH3-R-N727	CAAGCAGAAGACGGCATACGAGAT	ACTGATCG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	TA	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATACTGATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTTGGCGCGAAATTGTGG
	MSH3-R-N728	CAAGCAGAAGACGGCATACGAGAT	TAGCTGCA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	AT	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATTAGCTGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCATAGTTTGGCGCGAAATTGTGG
	MSH3-R-N729	CAAGCAGAAGACGGCATACGAGAT	GACGTCGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC	G	AGTTTGGCGCGAAATTGTGG	CAAGCAGAAGACGGCATACGAGATGACGTCGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCGAGTTTGGCGCGAAATTGTGG

Table 10.1. Nextera XT Index Kit v2 primers for the MSH3 repeat region.

10.7 PhIX reference sequence

>Illumina Enterobacteria phage phiX, complete genome

GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTTGATAAAGCAGGAATTACTACTGCTTGTTTAC
GAATTAAATCGAAGTGGACTGCTGGCGGAAAATGAGAAAATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTCGCCATCAACTAA
CGATTCTGTCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGGTTTAGATATGAGTCACATTTTGTTCAT
GGTAGAGATTCTCTTGTTGACATTTTAAAAGAGCGTGGATTACTATCTGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACT
GAACAATCCGTACGTTTCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATGATTTCGATTTTCTGACGAG
TAACAAAGTTTGGATTGCTACTGACCGCTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTCCTG
CTCCTGTTGAGTTTATTGCTGCCGTCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTACGGAAAAC
ATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGAATTGTTCGCGTTTACCTTGCGTGTACGCGCAGGAAACACTGACGTTCTTACTGACGCAGAAGA
AAACGTGCGTCAAAAATTACGTGCAGAAGGAGTGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTTGCGAGGTAC
TAAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGGGCTTCGGCCCCTTACTTGAGGATAAATTATGTCTAATA
TTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCATCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTACTCCGGTTATC
GCTGGCGACTCCTTCGAGATGGACGCCGTTGGCGCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGACTCTACTGTAGACATTTTTACTTTTTA
TGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTG
ACCATGCCGCTTTTCTTGGCACGATTAACCCTGATACCAATAAAATCCCTAAGCATTTGTTTCAGGGTTATTTGAATATCTATAACAACTATTTTAAAGCG
CCGTGGATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGCTCGTTATGGTTTCCGTTGCTGCCATCTCAAAAACATTTGGACTGC
TCCGCTTCCTCCTGAGACTGAGCTTTCTCGCCAAATGACGACTTCTACCACATCTATTGACATTATGGGTCTGCAAGCTGCTTATGCTAATTTGCATACTG
ACCAAGAACGTGATTACTTCATGCAGCGTTACCATGATGTTATTTCTTCATTTGGAGGTAAAACCTCTTATGACGCTGACAACCGTCCTTTACTTGTCATG
CGCTCTAATCTCTGGGCATCTGGCTATGATGTTGATGGAACTGACCAAACGTCGTTAGGCCAGTTTTCTGGTCGTGTTCAACAGACCTATAAACATTCTGT
GCCGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTTGTTCGTTTTCCGCCTACTGCGACTAAAGAGATTCAGTACCTTAACGCTAAAG
GTGCTTTGACTTATACCGATATTGCTGGCGACCCTGTTTTGTATGGCAACTTGCCGCCCGCGTGAAATTTCTATGAAGGATGTTTTCCGTTCTGGTGATTCG
TCTAAGAAGTTTAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCTTGAAGGCTTCCCATTCATTCAGGA
ACCGCCTTCTGGTGATTTGCAAGAACGCGTACTTATTCGCCACCATGATTATGACCAGTGTTTCCAGTCCGTTCAGTTGTTGCAGTGGAATAGTCAGGTTA
AATTTAATGTGACCGTTTATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTATAACGCCGAAGCGGTAA
AAATTTTAATTTTTGCCGCTGAGGGGTTGACCAAGCGAAGCGCGGTAGGTTTTCTGCTTAGGAGTTTAATCATGTTTCAGACTTTTATTTCTCGCCATAAT
TCAAACTTTTTTTCTGATAAGCTGGTTCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGTCAACGTTATATTT
TGATAGTTTGACGGTTAATGCTGGTAATGGTGGTTTTCTTCATTGCATTCAGATGGATACATCTGTCAACGCCGCTAATCAGGTTGTTTCTGTTGGTGCTG
ATATTGCTTTTGATGCCGACCCTAAATTTTTTGCCTGTTTGGTTCGCTTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCT
TTGGATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGTACGCCGGGCAATAATGTTTATGTTGGTTTCAT
GGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGTTTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCCAGCCACTTAAGTGAGGTGATT
TATGTTTGGTGCTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGTGGCGCCATGTCTAAATTGTTTGGAGGCGGTCAAAAAGCCGCCTCCGGTGGCATTC
AAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCTGGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCCTGATGAGGCC
GTCCCTAGTTTTGTTTCTGGTGCTATGGCTAAAGCTGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGATAAGTTGCTTGA
TTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTATCTTGCTGCTGCATTTCCTGAGCTTAATGCTTGGGAGCGTGCTG
GTGCTGATGCTTCCTCTGCTGGTATGGTTGACGCCGGATTTGAGAATCAAAAAGAGCTTACTAAAATGCAACTGGACAATCAGAAAGAGATTGCCGAGATG
CAAAATGAGACTCAAAAAGAGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGACCAGGTATATGCACAAAATGAGATGCTTGCTTATCA
ACAGAAGGAGTCTACTGCTCGCGTTGCGTCTATTATGGAAAACACCAATCTTTCCAAGCAACAGCAGGTTTCCGAGATTATGCGCCAAATGCTTACTCAAG
CTCAAACGGCTGGTCAGTATTTTACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTGACTTAGTTCATCAGCAAACGCAGAATCAG
CGGTATGGCTCTTCTCATATTGGCGCTACTGCAAAGGATATTTCTAATGTCGTCACTGATGCTGCTTCTGGTGTGGTTGATATTTTTCATGGTATTGATAA
AGCTGTTGCCGATACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTAATTTGTCTAGGAAATAACCGTCAGGATTGACACCCTCCC
AATTGTATGTTTTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTCGTTCTTATTACCCTTCTGAATGTCACGCTGATTATTTTGACTTTGAGCGTATC
GAGGCTCTTAAACCTGCTATTGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTTCCATAAGCAGATGGATAACCGCATCAAGCTCTT
GGAAGAGATTCTGTCTTTTCGTATGCAGGGCGTTGAGTTCGATAATGGTGATATGTATGTTGACGGCCATAAGGCTGCTTCTGACGTTCGTGATGAGTTTG
TATCTGTTACTGAGAAGTTAATGGATGAATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATATTAATAACACTATAGACCACCGCCCCGAAGGGGAC
GAAAAATGGTTTTTAGAGAACGAGAAGACGGTTACGCAGTTTTGCCGCAAGCTGGCTGCTGAACGCCCTCTTAAGGATATTCGCGATGAGTATAATTACCC
CAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATTGCTGGAGGCCTCCACTATGAAATCGCGTAGAGGCTTTGCTATTCAGCGTTTGATGAATGCAATGC
GACAGGCTCATGCTGATGGTTGGTTTATCGTTTTTGACACTCTCACGTTGGCTGACGACCGATTAGAGGCGTTTTATGATAATCCCAATGCTTTGCGTGAC
TATTTTCGTGATATTGGTCGTATGGTTCTTGCTGCCGAGGGTCGCAAGGCTAATGATTCACACGCCGACTGCTATCAGTATTTTTGTGTGCCTGAGTATGG
TACAGCTAATGGCCGTCTTCATTTCCATGCGGTGCACTTTATGCGGACACTTCCTACAGGTAGCGTTGACCCTAATTTTGGTCGTCGGGTACGCAATCGCC
GCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCATCGCAGTTCGCTACACGCAGGACGCTTTTTCACGTTCTGGTTGGTTGTGG
CCTGTTGATGCTAAAGGTGAGCCGCTTAAAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGGCTAAATACGTTAACAAAAAGTCAGATATGGACCTTGC
TGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACTTCCCAAGAAGCTGTTCAGAATCAGAATGAGCCGCAACTTCG
GGATGAAAATGCTCACAATGACAAATCTGTCCACGGAGTGCTTAATCCAACTTACCAAGCTGGGTTACGACGCGACGCCGTTCAACCAGATATTGAAGCAG
AACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGCGGCGCAACCTGTGACGACAAATCTGCTCAAATTTATGCGCGC
TTCGATAAAAATGATTGGCGTATCCAACCTGCA

10.8 MSH3 reference sequences

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGT
CTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGT
AGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCT
G

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGT
AGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCT
G

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGT
AGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCT
G

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG4_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG5_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCCGCCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG5_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGC
GCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGT
AGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCT
G

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCGCCTTTCCCGCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCGCTTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG4_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCGCTTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG5_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGC
GCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG5_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGC
GCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGT
AGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCT
G

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG4_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGC
GCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG5_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGC
GCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG5_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCA
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGC
GCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG4_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGC
GCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG5_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGC
GCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG5_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT0

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGC
GCCCCCAGCGCCCCCAGCGCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCG
GGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGT
AGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCT
G

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGT
AGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCT
G

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGT
AGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCT
G

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG4_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTT
TCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG5_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGC
TCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG5_CCCGCAGCG0_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGC
TCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGC
TCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG4_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG5_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG5_CCCGCAGCG1_CCCCCAGCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGC
GCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCG
GGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCC
GCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGAT
GATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
TCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
TCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG4_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG5_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG5_CCCGCAGCG0_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCG
GGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCA
GCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCT
CCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG2_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTT
CCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGC
GGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
TCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
TCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTGGGTAAGGCG
GGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG4_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCG
GGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG5_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCG
GGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG5_CCCGCAGCG1_CCCCCAGCG2_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGGCCGCAGCGGCCGCAGCGCCCGCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCT
GGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG1_GCCGCAGCG1_CCCGCAGCG1_GCCGCAGCG2_CCCGCAGCG1_CCCCCA
GCG1_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCGCAGCGGCCGCAGCGCCCGCAGCGGCCGCAGCGGCCGCAGC
GCCCGCAGCGCCCCCAGCGCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCT
GGGGGGGCGGGGCTTGTGGGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAGCG0_CCCCCAGCG5_CCCCCAGCT1

GAAATTGTGGCCGCCCCGCCCCCCTCGTCCCCATTTGTGCAGGCGAGGCCCCGCCCCCCCGCCCCGGCGCACGCAGGGTCGCGGCGTGCTCGCGCCCGC
AGACGCCTGGGAACTGCGGCCGCGGGCTCGCGCTCCTCGCCAGGCCCTGCCGCCGGGCTGCCATCCTTGCCCTGCCATGTCTCGCCGGAAGCCTGCGTC
GGGCGGCCTCGCTGCCTCCAGCTCAGCCCCTGCGAGGCAAGCGGTTTTGAGCCGATTCTTCCAGTCTACGGGAAGCCTGAAATCCACCTCCTCCTCCAC
AGGTGCAGCCGACCAGGTGGACCCTGGCGCTGCAGCGGCTGCAGCGGCCGCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGCGCCCCCAGC
GCCCCCAGCTCCCGCCTTCCCGCCCCAGCTGCCGCCGCACATAGTAGGTTCTGTCTGGGACTGGGCAGGGCCATCGGGGCTGGGGGGGCGGGGCTTGTG
GGTAAGGCGGGCGGAGGCGGGGACCCTCCGCCCGATGATAGGGCTG

10.9 MiSeq library quality control Galaxy workflow

Step 1: Input dataset

R1 fastq input

select at runtime

Step 2: Input dataset

IlluminaPHIXref

select at runtime

Step 3: Input dataset

R2 fastq input

select at runtime

Step 4: seqtk_sample

Input FASTA/Q file

Output dataset 'output' from step 1

RNG seed

4

Subsample (decimal fraction or number)

200000.0

Step 5: seqtk_sample

Input FASTA/Q file

Output dataset 'output' from step 3

RNG seed

4

Subsample (decimal fraction or number)

200000.0

Step 6: FastQC

Short read data from your current history

Output dataset 'default' from step 4

Contaminant list

select at runtime

Submodule and Limit specifing file

select at runtime

Step 7: Cutadapt

Single-end or Paired-end reads?

Single-end

FASTQ/A file

Output dataset 'default' from step 4

Read 1 Options:

3' (End) Adapters

3' (End) Adapters 1

Source

Enter custom sequence

Enter custom 3' adapter name (Optional)

Illumina Seq primer binding site

Enter custom 3' adapter sequence

GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

5' (Front) Adapters

5' or 3' (Anywhere) Adapters

Cut bases from reads before adapter trimming

0

Adapter Options:

Maximum error rate

0.39

Do not allow indels (Use ONLY with anchored 5' (front) adapters).

False

Match times

1

Minimum overlap length

3

Match Read Wildcards

False

Filter Options:

Discard Trimmed Reads

False

Discard Untrimmed Reads

False

Minimum length

0

Maximum length

0

Do not trim adapters

False

Mask Adapters

False

Max N

Not available.

Pair filter

any

Read Modification Options:

Quality cutoff

0

NextSeq trimming

0

Trim Ns

False

Prefix

Empty.

Suffix

Empty.

Strip suffix

Empty.

Length

0

Length Tag

Empty.

Output Options:

Report

False

Info File

False

Rest of Read

False

Wildcard File

False
Too Short Reads
False
Too Long Reads
False
Untrimmed Reads
False

Step 8: Cutadapt

Single-end or Paired-end reads?
Single-end
FASTQ/A file
Output dataset 'default' from step 5
Read 1 Options:
3' (End) Adapters
3' (End) Adapters 1
Source
Enter custom sequence
Enter custom 3' adapter name (Optional)
Illumina Forward Seq primer binding site
Enter custom 3' adapter sequence
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

5' (Front) Adapters
5' or 3' (Anywhere) Adapters
Cut bases from reads before adapter trimming
0

Adapter Options:
Maximum error rate
0.1
Do not allow indels (Use ONLY with anchored 5' (front) adapters).
False
Match times
1
Minimum overlap length
3
Match Read Wildcards
False

Filter Options:
Discard Trimmed Reads
False
Discard Untrimmed Reads
False
Minimum length
0
Maximum length
0
Do not trim adapters
False
Mask Adapters
False
Max N
Not available.
Pair filter
any

Read Modification Options:
Quality cutoff
0
NextSeq trimming
0
Trim Ns
False
Prefix
Empty.
Suffix
Empty.
Strip suffix
Empty.
Length
0
Length Tag
Empty.

Output Options:
Report
False
Info File
False
Rest of Read
False
Wildcard File
False
Too Short Reads
False
Too Long Reads
False
Untrimmed Reads
False

Step 9: FastQC

Short read data from your current history
Output dataset 'default' from step 5
Contaminant list
select at runtime
Submodule and Limit specifing file
select at runtime

Step 10: FastQC

Short read data from your current history
Output dataset 'out1' from step 7
Contaminant list
select at runtime
Submodule and Limit specifing file
select at runtime

Step 11: Map with BWA-MEM

Will you select a reference genome from your history or use a built-in index?
Use a genome from history and build index
Use the following dataset as the reference sequence
Output dataset 'output' from step 2
Algorithm for constructing the BWT index
Auto. Let BWA decide the best algorithm to use

Single or Paired-end reads
Single
Select fastq dataset
Output dataset 'out1' from step 7
Set read groups information?
Do not set
Select analysis mode
1.Simple Illumina mode

Step 12: Map with BWA-MEM

Will you select a reference genome from your history or use a built-in index?
Use a genome from history and build index
Use the following dataset as the reference sequence
Output dataset 'output' from step 2
Algorithm for constructing the BWT index
Auto. Let BWA decide the best algorithm to use
Single or Paired-end reads
Single
Select fastq dataset
Output dataset 'out1' from step 8
Set read groups information?
Do not set
Select analysis mode
1.Simple Illumina mode

Step 13: FastQC

Short read data from your current history
Output dataset 'out1' from step 8
Contaminant list
select at runtime
Submodule and Limit specifing file
select at runtime

Step 14: BAM-to-SAM

BAM File to Convert
Output dataset 'bam_output' from step 11
Header options
Include header in SAM output (-h)

Step 15: BAM-to-SAM

BAM File to Convert
Output dataset 'bam_output' from step 12
Header options
Include header in SAM output (-h)

Step 16: Filter

Filter
Output dataset 'output1' from step 14
With following condition
c5>0
Number of header lines to skip
3

Step 17: Filter

Filter
Output dataset 'output1' from step 15
With following condition
c5>0
Number of header lines to skip
3

Step 18: SAM-to-BAM

Choose the source for the reference genome
Use a genome from the history
SAM file to convert
Output dataset 'out_file1' from step 16
Using reference file
Output dataset 'output' from step 2

Step 19: SAM-to-BAM

Choose the source for the reference genome
Use a genome from the history
SAM file to convert
Output dataset 'out_file1' from step 17
Using reference file
Output dataset 'output' from step 2

Step 20: Convert from BAM to FastQ

Convert the following BAM file to FASTQ
Output dataset 'output1' from step 18
Create FASTQ based on the mate info in the BAM R2 and Q2 tags.
False
FASTQ for second end. Used if BAM contains paired-end data. BAM should be sorted by query name if creating paired FASTQ with this option.
False

Step 21: Convert from BAM to FastQ

Convert the following BAM file to FASTQ
Output dataset 'output1' from step 19
Create FASTQ based on the mate info in the BAM R2 and Q2 tags.
False
FASTQ for second end. Used if BAM contains paired-end data. BAM should be sorted by query name if creating paired FASTQ with this option.
False

Step 22: Cutadapt

Single-end or Paired-end reads?
Single-end
FASTQ/A file
Output dataset 'output' from step 20
Read 1 Options:
3' (End) Adapters
5' (Front) Adapters
5' or 3' (Anywhere) Adapters
Cut bases from reads before adapter trimming
0

Adapter Options:
Maximum error rate
0.1
Do not allow indels (Use ONLY with anchored 5' (front) adapters).

False
Match times
1
Minimum overlap length
3
Match Read Wildcards
False

Filter Options:
Discard Trimmed Reads
False
Discard Untrimmed Reads
False
Minimum length
300
Maximum length
0
Do not trim adapters
False
Mask Adapters
False
Max N
Not available.
Pair filter
any

Read Modification Options:
Quality cutoff
0
NextSeq trimming
0
Trim Ns
False
Prefix
Empty.
Suffix
Empty.
Strip suffix
Empty.
Length
0
Length Tag
Empty.

Output Options:
Report
False
Info File
False
Rest of Read
False
Wildcard File
False
Too Short Reads
False
Too Long Reads
False
Untrimmed Reads
False

Step 23: Cutadapt
Single-end or Paired-end reads?
Single-end
FASTQ/A file
Output dataset 'output' from step 20
Read 1 Options:
3' (End) Adapters
5' (Front) Adapters
5' or 3' (Anywhere) Adapters
Cut bases from reads before adapter trimming
0

Adapter Options:
Maximum error rate
0.1
Do not allow indels (Use ONLY with anchored 5' (front) adapters).
False
Match times
1
Minimum overlap length
3
Match Read Wildcards
False

Filter Options:
Discard Trimmed Reads
False
Discard Untrimmed Reads
False
Minimum length
390
Maximum length
0
Do not trim adapters
False
Mask Adapters
False
Max N
Not available.
Pair filter
any

Read Modification Options:
Quality cutoff
0
NextSeq trimming
0
Trim Ns
False
Prefix
Empty.
Suffix
Empty.
Strip suffix
Empty.

Length
0
Length Tag
Empty.

Output Options:
Report
False
Info File
False
Rest of Read
False
Wildcard File
False
Too Short Reads
False
Too Long Reads
False
Untrimmed Reads
False

Step 24: Cutadapt
Single-end or Paired-end reads?
Single-end
FASTQ/A file
Output dataset 'output' from step 21
Read 1 Options:
3' (End) Adapters
5' (Front) Adapters
5' or 3' (Anywhere) Adapters
Cut bases from reads before adapter trimming
0

Adapter Options:
Maximum error rate
0.1
Do not allow indels (Use ONLY with anchored 5' (front) adapters).
False
Match times
1
Minimum overlap length
3
Match Read Wildcards
False

Filter Options:
Discard Trimmed Reads
False
Discard Untrimmed Reads
False
Minimum length
100
Maximum length
0
Do not trim adapters
False
Mask Adapters
False
Max N
Not available.
Pair filter
any

Read Modification Options:
Quality cutoff
0
NextSeq trimming
0
Trim Ns
False
Prefix
Empty.
Suffix
Empty.
Strip suffix
Empty.
Length
0
Length Tag
Empty.

Output Options:
Report
False
Info File
False
Rest of Read
False
Wildcard File
False
Too Short Reads
False
Too Long Reads
False
Untrimmed Reads
False

Step 25: Cutadapt
Single-end or Paired-end reads?
Single-end
FASTQ/A file
Output dataset 'output' from step 21
Read 1 Options:
3' (End) Adapters
5' (Front) Adapters
5' or 3' (Anywhere) Adapters
Cut bases from reads before adapter trimming
0

Adapter Options:
Maximum error rate
0.1
Do not allow indels (Use ONLY with anchored 5' (front) adapters).
False
Match times
1

Minimum overlap length
3
Match Read Wildcards
False

Filter Options:
Discard Trimmed Reads
False
Discard Untrimmed Reads
False
Minimum length
190
Maximum length
0
Do not trim adapters
False
Mask Adapters
False
Max N
Not available.
Pair filter
any

Read Modification Options:
Quality cutoff
0
NextSeq trimming
0
Trim Ns
False
Prefix
Empty.
Suffix
Empty.
Strip suffix
Empty.
Length
0
Length Tag
Empty.

Output Options:
Report
False
Info File
False
Rest of Read
False
Wildcard File
False
Too Short Reads
False
Too Long Reads
False
Untrimmed Reads
False

Step 26: Trim

Input dataset
Output dataset 'out1' from step 22
Trim this column only
0
Trim from the beginning up to this position
1
Remove everything from this position to the end
300
Is input dataset in FASTQ format?
Yes
Ignore lines beginning with these characters
Nothing selected.

Step 27: FastQC

Short read data from your current history
Output dataset 'out1' from step 23
Contaminant list
select at runtime
Submodule and Limit specifing file
select at runtime

Step 28: Raspberry QC

Short read data from your current history
Output dataset 'out1' from step 23

Step 29: Trim

Input dataset
Output dataset 'out1' from step 24
Trim this column only
0
Trim from the beginning up to this position
1
Remove everything from this position to the end
100
Is input dataset in FASTQ format?
Yes
Ignore lines beginning with these characters
Nothing selected.

Step 30: FastQC

Short read data from your current history
Output dataset 'out1' from step 25
Contaminant list
select at runtime
Submodule and Limit specifing file
select at runtime

Step 31: Raspberry QC

Short read data from your current history
Output dataset 'out1' from step 25

Step 32: Raspberry QC

Short read data from your current history
Output dataset 'out_file1' from step 26

Step 33: Raspberry QC

Short read data from your current history
Output dataset 'out_file1' from step 29

Step 34: Concatenate datasets

Concatenate Dataset

Output dataset 'output' from step 32

Datasets

Dataset 1

Select

Output dataset 'output' from step 28

Step 35: Concatenate datasets

Concatenate Dataset

Output dataset 'output' from step 31

Datasets

Dataset 1

Select

Output dataset 'output' from step 33

10.10 MSH3 repeat genotyping Galaxy workflow

Step 1: Input dataset

R1 read file

select at runtime

Step 2: Input dataset

R2 read file

select at runtime

Step 3: Input dataset

Reference file

select at runtime

Step 4: Pear

Dataset type

Paired-end

Name of file that contains the forward paired-end reads

Output dataset 'output' from step 1

Name of file that contains the reverse paired-end reads

Output dataset 'output' from step 2

Specify a p-value for the statistical test

0.01

Minimum overlap size

10

Maximum possible length of the assembled sequences

0

Minimum possible length of the assembled sequences

50

Minimum length of reads after trimming the low quality part

1

Quality score threshold for trimming the low quality part of a read

0

Maximal proportion of uncalled bases in a read

1.0

Specify the upper bound for the resulting quality score

40

Type of statistical test

Given the minimum allowed overlap, test using the highest OES (1)

Disable empirical base frequencies

False

Use N base if uncertain

False

Scoring method

Assembly score (AS) use +1 for match and -1 for mismatch multiplied by base quality scores

Output files

Assembled reads

Step 5: Cutadapt

Single-end or Paired-end reads?

Single-end

FASTQ/A file

Output dataset 'assembled_reads' from step 4

Read 1 Options:

3' (End) Adapters

5' (Front) Adapters

5' (Front) Adapters 1

Source

Enter custom sequence

Enter custom 5' adapter name (Optional)

10 bp of forward gene specific primer

Enter custom 5' adapter sequence

CTTCCTCCTC

5' or 3' (Anywhere) Adapters

Cut bases from reads before adapter trimming

0

Adapter Options:

Maximum error rate

0.1

Do not allow indels (Use ONLY with anchored 5' (front) adapters).

False

Match times

1

Minimum overlap length

3

Match Read Wildcards

False

Filter Options:

Discard Trimmed Reads

False

Discard Untrimmed Reads

True

Minimum length

0

Maximum length

0

Do not trim adapters

False

Mask Adapters

False

Max N

Not available.

Pair filter

any

Read Modification Options:

Quality cutoff

0

NextSeq trimming

0

Trim Ns

False

Prefix

Empty.

Suffix

Empty.

Strip suffix

Empty.
Length
0
Length Tag
Empty.

Output Options:
Report
False
Info File
False
Rest of Read
False
Wildcard File
False
Too Short Reads
False
Too Long Reads
False
Untrimmed Reads
False

Step 6: Cutadapt
Single-end or Paired-end reads?
Single-end
FASTQ/A file
Output dataset 'out1' from step 5
Read 1 Options:
3' (End) Adapters
3' (End) Adapters 1
Source
Enter custom sequence
Enter custom 3' adapter name (Optional)
last 10bp of reverse gene specific primer sequence
Enter custom 3' adapter sequence
GCGCCAAACT

5' (Front) Adapters
5' or 3' (Anywhere) Adapters
Cut bases from reads before adapter trimming
0

Adapter Options:
Maximum error rate
0.1
Do not allow indels (Use ONLY with anchored 5' (front) adapters).
False
Match times
1
Minimum overlap length
3
Match Read Wildcards
False

Filter Options:
Discard Trimmed Reads
False
Discard Untrimmed Reads
True
Minimum length
0
Maximum length
0
Do not trim adapters
False
Mask Adapters
False
Max N
Not available.
Pair filter
any

Read Modification Options:
Quality cutoff
0
NextSeq trimming
0
Trim Ns
False
Prefix
Empty.
Suffix
Empty.
Strip suffix
Empty.
Length
0
Length Tag
Empty.

Output Options:
Report
False
Info File
False
Rest of Read
False
Wildcard File
False
Too Short Reads
False
Too Long Reads
False
Untrimmed Reads
False

Step 7: FastQC
Short read data from your current history
Output dataset 'out1' from step 5
Contaminant list
select at runtime
Submodule and Limit specifing file
select at runtime

Step 8: Map with BWA-MEM

Will you select a reference genome from your history or use a built-in index?
Use a genome from history and build index
Use the following dataset as the reference sequence
Output dataset 'output' from step 3
Algorithm for constructing the BWT index
Auto. Let BWA decide the best algorithm to use
Single or Paired-end reads
Single
Select fastq dataset
Output dataset 'out1' from step 6
Set read groups information?
Do not set
Select analysis mode
5.Full list of options
Set algorithmic options?
Set
Minimum seed length
19
Band width for banded alignment
100
Off-diagonal X-dropoff
100
Look for internal seeds inside a seed longer than -k * THIS VALUE
1.5
Seed occurrence for the 3rd round seeding
20
Skip seeds with more than that many occurrences
500
Drop chains shorter than this fraction of the longest overlapping chain
0.5
Discard a chain if seeded bases shorter than THIS VALUE
0
Perform at most this many rounds of mate rescues for each read
50
Skip mate rescue
False
Skip pairing; mate rescue performed unless -S also in use
False
Discard full-length exact matches
False
Set scoring options?
Set
Score for a sequence match
1
Penalty for a mismatch
5
Gap open penalties for deletions and insertions
6,6
Gap extension penalties; a gap of size k cost '-O + -E*k'. If two numbers are specified, the first is the penalty of extending a deletion and the second for extending an insertion
1,1
Penalties for 5'-end and 3'-end clipping
5,5
Penalty for an unpaired read pair
17
Set input/output options
Do not set

Step 9: FastQC

Short read data from your current history
Output dataset 'out1' from step 6
Contaminant list
select at runtime
Submodule and Limit specifing file
select at runtime

Step 10: BAM-to-SAM

BAM File to Convert
Output dataset 'bam_output' from step 8
Header options
Include header in SAM output (-h)

Step 11: Filter

Filter
Output dataset 'output1' from step 10
With following condition
c5>0
Number of header lines to skip
84

Step 12: Repeat Counter

Short read data from your current history
Output dataset 'out_file1' from step 11

Step 13: Convert

Convert all
Commas
in Dataset
Output dataset 'repeat_counts.csv' from step 12
Strip leading and trailing whitespaces
True
Condense consecutive delimiters in one TAB
True

Step 14: Sort

Sort Dataset
Output dataset 'out_file1' from step 13
on column
2
with flavor
Numerical sort
everything in
Descending order
Column selections

10.11 MSH3 variant calling Galaxy workflow

Step 1: Input dataset

BAM input

select at runtime

Step 2: Input dataset

Reference file

select at runtime

Step 3: Naive Variant Caller (NVC)

Choose the source for the reference list

History

BAM files

BAM file 1

BAM file

Output dataset 'output' from step 1

Using reference file

Output dataset 'output' from step 2

Restrict to regions

Restrict to regions 1

Chromosome

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG1_CCCGCAG
CG0_CCCCCAGCG1_CCCCCAGCT0

Start

1

End

483

Restrict to regions 2

Chromosome

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAG
CG0_CCCCCAGCG1_CCCCCAGCT1

Start

1

End

501

Restrict to regions 3

Chromosome

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAG
CG0_CCCCCAGCG0_CCCCCAGCT1

Start

1

End

501

Restrict to regions 4

Chromosome

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAG
CG0_CCCCCAGCG1_CCCCCAGCT1

Start

1

End

510

Restrict to regions 5

Chromosome

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAG
CG1_CCCCCAGCG1_CCCCCAGCT1*_CCCCCAGCT1

Start

1

End

510

Restrict to regions 6

Chromosome

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAG
CG0_CCCCCAGCG1_CCCCCAGCT1

Start

1

End

510

Restrict to regions 7

Chromosome

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAG
CG1_CCCCCAGCG1_CCCCCAGCT1

Start

1

End

519

Restrict to regions 8

Chromosome

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG2_CCCGCAG
CG0_CCCCCAGCG2_CCCCCAGCT1

Start

1

End

519

Restrict to regions 9

Chromosome

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAG
CG0_CCCCCAGCG2_CCCCCAGCT1

Start

1

End

519

Restrict to regions 10

Chromosome

>MSH3_reference_GCTGCAGCG1_GCCGCAGCG3_CCCGCAG
CG1_CCCCCAGCG1_CCCCCAGCT1

Start

1

End

519

Restrict to regions 11

Chromosome

>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAG
CG0_CCCCCAGCG1_CCCCCAGCT1

Start

1
End
537

Restrict to regions 12
Chromosome
>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAG
CG1_CCCCCAGCG1_CCCCCAGCT1
Start
Not available.
End
Not available.

Restrict to regions 13
Chromosome
>MSH3_reference_GCTGCAGCG1_GCCGCAGCG4_CCCGCAG
CG1_CCCCCAGCG1_CCCCCAGCT1
Start
Not available.
End
Not available.

Restrict to regions 14
Chromosome
>MSH3_reference_GCTGCAGCG2_GCCGCAGCG1_CCCGCAG
CG0_CCCCCAGCG5_CCCCCAGCT1
Start
Not available.
End
Not available.

Restrict to regions 15
Chromosome
>MSH3_reference_GCTGCAGCG2_GCCGCAGCG3_CCCGCAG
CG1_CCCCCAGCG2_CCCCCAGCT1
Start
Not available.
End
Not available.

Restrict to regions by files
Minimum number of reads needed to consider a REF/ALT
100
Minimum base quality
20
Minimum mapping quality
1
Ploidy
2
Only write out positions with possible alternate alleles
False
Report counts by strand
True
Show Advanced Options
Hide Advanced Options

Step 4: VCFfilter:
Specify filterting expression
-f "AF > 0.4"
VCF dataset to filter
Output dataset 'output_vcf' from step 3

Filtering for allele freequncy

Step 5: VCFtoTab-delimited:
Select VCF dataset to convert
Output dataset 'out_file1' from step 4
Report data per sample
True
Fill empty fields with
Nothing

10.12 Base-wise conservation scores across the MSH3 9bp tandem repeat region

Table 10.2. Base-wise conservation scores across the MSH3 exon 1 9bp tandem repeat region.
PhastCons and PhyloP generated values are displayed for bases on chromosome 5 at positions 80654862 to 80654969. Tandem repeat region is shown in bold. Generated on UCSC.

Chromosomal position	PhastCons	PhyloP
80654862	0	-0.01
80654863	0	-0.34
80654864	0	0.07
80654865	0	-0.17
80654866	0	0.23
80654867	0	-0.42
80654868	0	-3.83
80654869	0	-0.09
80654870	0	-0.09
80654871	0	0.07
80654872	0	-0.58
80654873	0	-0.17
80654874	0	-2.53
80654875	0.45	3.24
80654876	0	-0.74
80654877	0	0.23
80654878	**0**	**-0.09**
80654879	**0.02**	**3.4**
80654880	**0**	**-0.82**
80654881	**0**	**1.04**
80654882	**0**	**0.07**
80654883	**0**	**-0.42**
80654884	**0**	**-0.66**
80654885	**0**	**0.64**
80654886	**0.01**	**0.07**
80654887	**0.01**	**0.07**
80654888	**0.01**	**0.07**
80654889	**0.01**	**-0.01**
80654890	**0.01**	**-0.01**
80654891	**0.01**	**-0.01**
80654892	**0.01**	**-0.01**
80654893	**0.01**	**-0.01**
80654894	**0.01**	**-0.01**
80654895	**0.01**	**-0.01**
80654896	**0**	**-0.01**
80654897	**0**	**-0.01**
80654898	**0**	**-0.42**
80654899	**0**	**-0.01**
80654900	**0**	**-0.01**
80654901	**0**	**-0.01**
80654902	**0**	**-0.01**
80654903	**0**	**-0.01**
80654904	**0**	**-0.01**
80654905	**0**	**-0.01**
80654906	**0**	**-0.01**
80654907	**0**	**-1.31**
80654908	**0.01**	**0.07**
80654909	**0.04**	**0.39**
80654910	**0.09**	**-0.01**
80654911	**0.92**	**1.77**
80654912	**0.93**	**1.13**
80654913	**0.93**	**0.31**
80654914	**0.94**	**1.13**
80654915	**0.97**	**2.91**
80654916	**0.97**	**0.39**
80654917	**0.97**	**1.37**
80654918	**0.96**	**-0.01**
80654919	**0.97**	**0.39**
80654920	**1**	**2.02**

80654921	**0.99**	**0.8**
80654922	**0.96**	**2.1**
80654923	**0.85**	**-0.01**
80654924	**0.75**	**-0.01**
80654925	**0.64**	**-0.01**
80654926	**0.54**	**-0.01**
80654927	**0.44**	**-0.01**
80654928	**0.34**	**-0.01**
80654929	**0.23**	**-0.01**
80654930	**0.13**	**-0.01**
80654931	**0.02**	**-2.37**
80654932	0	-1.39
80654933	0	0.31
80654934	0	0.15
80654935	0	-0.58
80654936	0	0.39
80654937	0	-0.5
80654938	0	-0.34
80654939	0	-1.63
80654940	0	-0.5
80654941	0	-0.01
80654942	0	-1.07
80654943	0	-0.01
80654944	0	-0.74
80654945	0	0.07
80654946	0	-0.5
80654947	0	-0.9
80654948	0	-0.99
80654949	0	0.07
80654950	0	-1.31
80654951	0	-0.34
80654952	0	0.31
80654953	0	-0.25
80654954	0	-0.99
80654955	0	-0.17
80654956	0	0.23
80654957	0	-0.25
80654958	0	-0.25
80654959	0	0.15
80654960	0	-0.82
80654961	0	-1.23
80654962	0	-0.25
80654963	0	-0.34
80654964	0	0.48
80654965	0	0.56
80654966	0	-0.01
80654967	0	0.39
80654968	0	-0.09
80654969	0	0.72

10.13 MSH3 and DHFR expression quantitative trait loci (eQTL)

Table 10.3. MSH3 and DHFR expression quantitative trait loci associated with phenotypes in HD and DM1.
Data from GTex. NES – normalised effect size, ref – reference allele.

SNP Id	location	ref	minor allele	Gene Symbol	p	NES	Tissue
rs151182735	5:80654571	C	-	DHFR	7.80E-92	-1	Muscle - Skeletal
rs151182735	5:80654571	C	-	DHFR	5.00E-50	-0.62	Adipose - Subcutaneous
rs151182735	5:80654571	C	-	DHFR	3.20E-49	-0.65	Whole Blood
rs151182735	5:80654571	C	-	DHFR	8.10E-45	-0.86	Testis
rs151182735	5:80654571	C	-	DHFR	3.40E-39	-0.56	Adipose - Visceral (Omentum)
rs151182735	5:80654571	C	-	DHFR	3.10E-34	-0.78	Heart - Left Ventricle
rs151182735	5:80654571	C	-	DHFR	3.30E-32	-0.3	Cells - Transformed fibroblasts
rs151182735	5:80654571	C	-	DHFR	2.40E-28	-0.8	Colon - Sigmoid
rs151182735	5:80654571	C	-	DHFR	4.80E-28	-0.69	Esophagus - Muscularis
rs151182735	5:80654571	C	-	DHFR	5.90E-27	-0.76	Heart - Atrial Appendage
rs151182735	5:80654571	C	-	DHFR	1.40E-23	-0.5	Cells - EBV-transformed lymphocytes
rs151182735	5:80654571	C	-	DHFR	1.20E-22	-0.29	Esophagus - Mucosa
rs151182735	5:80654571	C	-	DHFR	3.10E-20	-0.42	Colon - Transverse
rs151182735	5:80654571	C	-	DHFR	4.90E-20	-0.95	Brain - Cortex
rs151182735	5:80654571	C	-	DHFR	3.10E-19	-0.32	Lung
rs151182735	5:80654571	C	-	DHFR	1.60E-18	-0.42	Breast - Mammary Tissue
rs151182735	5:80654571	C	-	DHFR	2.40E-18	-0.93	Brain - Anterior cingulate cortex (BA24)
rs151182735	5:80654571	C	-	DHFR	5.10E-16	-0.31	Skin - Not Sun Exposed (Suprapubic)
rs151182735	5:80654571	C	-	DHFR	1.00E-15	-0.42	Artery - Tibial
rs151182735	5:80654571	C	-	DHFR	1.90E-15	-0.3	Skin - Sun Exposed (Lower leg)
rs151182735	5:80654571	C	-	DHFR	2.90E-15	-0.64	Brain - Nucleus accumbens (basal ganglia)
rs151182735	5:80654571	C	-	DHFR	2.00E-14	-0.85	Brain - Frontal Cortex (BA9)
rs151182735	5:80654571	C	-	DHFR	3.50E-14	-0.43	Nerve - Tibial
rs151182735	5:80654571	C	-	DHFR	5.00E-14	-0.53	Spleen
rs151182735	5:80654571	C	-	DHFR	1.50E-13	-0.65	Brain - Caudate (basal ganglia)
rs151182735	5:80654571	C	-	DHFR	3.60E-13	-0.76	Brain - Cerebellum
rs151182735	5:80654571	C	-	DHFR	4.90E-13	-0.58	Esophagus - Gastroesophageal Junction
rs151182735	5:80654571	C	-	DHFR	1.60E-12	-0.64	Brain - Hypothalamus
rs151182735	5:80654571	C	-	DHFR	3.10E-12	-0.76	Brain - Cerebellar Hemisphere
rs151182735	5:80654571	C	-	DHFR	5.00E-12	-0.62	Brain - Hippocampus
rs151182735	5:80654571	C	-	DHFR	2.10E-11	-0.61	Brain - Amygdala
rs151182735	5:80654571	C	-	DHFR	5.40E-10	-0.57	Adrenal Gland
rs151182735	5:80654571	C	-	DHFR	4.70E-09	-0.49	Pituitary
rs151182735	5:80654571	C	-	DHFR	1.60E-08	-0.71	Brain - Spinal cord (cervical c-1)
rs151182735	5:80654571	C	-	DHFR	2.00E-08	-0.59	Brain - Substantia nigra
rs151182735	5:80654571	C	-	DHFR	3.30E-07	-0.4	Artery - Aorta
rs151182735	5:80654571	C	-	DHFR	3.30E-07	-0.47	Artery - Coronary
rs151182735	5:80654571	C	-	DHFR	3.80E-07	-0.52	Prostate
rs151182735	5:80654571	C	-	DHFR	1.10E-05	-0.4	Brain - Putamen (basal ganglia)
rs151182735	5:80654571	C	-	MSH3	2.10E-28	-0.49	Thyroid
rs151182735	5:80654571	C	-	MSH3	7.30E-25	-0.42	Skin - Sun Exposed (Lower leg)
rs151182735	5:80654571	C	-	MSH3	1.60E-24	-0.49	Artery - Tibial
rs151182735	5:80654571	C	-	MSH3	3.30E-23	-0.45	Skin - Not Sun Exposed (Suprapubic)
rs151182735	5:80654571	C	-	MSH3	3.00E-21	-0.49	Nerve - Tibial
rs151182735	5:80654571	C	-	MSH3	5.70E-21	-0.4	Adipose - Subcutaneous
rs151182735	5:80654571	C	-	MSH3	2.00E-18	-0.4	Whole Blood
rs151182735	5:80654571	C	-	MSH3	5.80E-16	-0.55	Cells - Transformed fibroblasts
rs151182735	5:80654571	C	-	MSH3	1.30E-14	-0.41	Heart - Atrial Appendage
rs151182735	5:80654571	C	-	MSH3	2.80E-14	-0.51	Artery - Aorta
rs151182735	5:80654571	C	-	MSH3	1.40E-10	-0.32	Adipose - Visceral (Omentum)
rs151182735	5:80654571	C	-	MSH3	2.10E-10	-0.44	Pancreas
rs151182735	5:80654571	C	-	MSH3	3.70E-09	-0.36	Colon - Sigmoid
rs151182735	5:80654571	C	-	MSH3	3.10E-08	-0.42	Adrenal Gland
rs151182735	5:80654571	C	-	MSH3	7.50E-08	-0.35	Breast - Mammary Tissue
rs151182735	5:80654571	C	-	MSH3	1.50E-07	-0.42	Pituitary
rs151182735	5:80654571	C	-	MSH3	2.20E-07	-0.3	Colon - Transverse
rs151182735	5:80654571	C	-	MSH3	6.80E-07	-0.26	Esophagus - Muscularis
rs151182735	5:80654571	C	-	MSH3	7.00E-07	-0.34	Esophagus - Gastroesophageal Junction
rs151182735	5:80654571	C	-	MSH3	7.90E-06	-0.18	Esophagus - Mucosa
rs10168	5:80654584	C	T	DHFR	8.90E-92	-1	Muscle - Skeletal
rs10168	5:80654584	C	T	DHFR	5.00E-50	-0.62	Adipose - Subcutaneous
rs10168	5:80654584	C	T	DHFR	2.50E-49	-0.65	Whole Blood
rs10168	5:80654584	C	T	DHFR	5.70E-44	-0.85	Testis
rs10168	5:80654584	C	T	DHFR	1.80E-39	-0.56	Adipose - Visceral (Omentum)
rs10168	5:80654584	C	T	DHFR	1.40E-34	-0.78	Heart - Left Ventricle
rs10168	5:80654584	C	T	DHFR	2.70E-32	-0.3	Cells - Transformed fibroblasts
rs10168	5:80654584	C	T	DHFR	2.40E-28	-0.8	Colon - Sigmoid
rs10168	5:80654584	C	T	DHFR	4.80E-28	-0.69	Esophagus - Muscularis
rs10168	5:80654584	C	T	DHFR	2.60E-27	-0.76	Heart - Atrial Appendage
rs10168	5:80654584	C	T	DHFR	1.70E-23	-0.5	Cells - EBV-transformed lymphocytes
rs10168	5:80654584	C	T	DHFR	1.20E-22	-0.29	Esophagus - Mucosa
rs10168	5:80654584	C	T	DHFR	2.30E-20	-0.42	Colon - Transverse
rs10168	5:80654584	C	T	DHFR	4.90E-20	-0.95	Brain - Cortex
rs10168	5:80654584	C	T	DHFR	3.10E-19	-0.32	Lung

rs10168	5:80654584	C	T	DHFR	1.20E-18	-0.42	Breast - Mammary Tissue
rs10168	5:80654584	C	T	DHFR	2.40E-18	-0.93	Brain - Anterior cingulate cortex (BA24)
rs10168	5:80654584	C	T	DHFR	5.10E-16	-0.31	Skin - Not Sun Exposed (Suprapubic)
rs10168	5:80654584	C	T	DHFR	1.00E-15	-0.42	Artery - Tibial
rs10168	5:80654584	C	T	DHFR	1.90E-15	-0.3	Skin - Sun Exposed (Lower leg)
rs10168	5:80654584	C	T	DHFR	2.90E-15	-0.64	Brain - Nucleus accumbens (basal ganglia)
rs10168	5:80654584	C	T	DHFR	2.00E-14	-0.85	Brain - Frontal Cortex (BA9)
rs10168	5:80654584	C	T	DHFR	3.50E-14	-0.43	Nerve - Tibial
rs10168	5:80654584	C	T	DHFR	5.00E-14	-0.53	Spleen
rs10168	5:80654584	C	T	DHFR	1.50E-13	-0.65	Brain - Caudate (basal ganglia)
rs10168	5:80654584	C	T	DHFR	3.60E-13	-0.76	Brain - Cerebellum
rs10168	5:80654584	C	T	DHFR	6.00E-13	-0.58	Esophagus - Gastroesophageal Junction
rs10168	5:80654584	C	T	DHFR	1.60E-12	-0.64	Brain - Hypothalamus
rs10168	5:80654584	C	T	DHFR	3.10E-12	-0.76	Brain - Cerebellar Hemisphere
rs10168	5:80654584	C	T	DHFR	5.00E-12	-0.62	Brain - Hippocampus
rs10168	5:80654584	C	T	DHFR	2.10E-11	-0.61	Brain - Amygdala
rs10168	5:80654584	C	T	DHFR	5.40E-10	-0.57	Adrenal Gland
rs10168	5:80654584	C	T	DHFR	4.70E-09	-0.49	Pituitary
rs10168	5:80654584	C	T	DHFR	1.60E-08	-0.71	Brain - Spinal cord (cervical c-1)
rs10168	5:80654584	C	T	DHFR	2.00E-08	-0.59	Brain - Substantia nigra
rs10168	5:80654584	C	T	DHFR	2.20E-07	-0.47	Artery - Coronary
rs10168	5:80654584	C	T	DHFR	3.30E-07	-0.4	Artery - Aorta
rs10168	5:80654584	C	T	DHFR	3.80E-07	-0.52	Prostate
rs10168	5:80654584	C	T	DHFR	1.10E-05	-0.4	Brain - Putamen (basal ganglia)
rs10168	5:80654584	C	T	MSH3	3.80E-28	-0.48	Thyroid
rs10168	5:80654584	C	T	MSH3	7.30E-25	-0.42	Skin - Sun Exposed (Lower leg)
rs10168	5:80654584	C	T	MSH3	1.60E-24	-0.49	Artery - Tibial
rs10168	5:80654584	C	T	MSH3	3.30E-23	-0.45	Skin - Not Sun Exposed (Suprapubic)
rs10168	5:80654584	C	T	MSH3	3.00E-21	-0.49	Nerve - Tibial
rs10168	5:80654584	C	T	MSH3	5.70E-21	-0.4	Adipose - Subcutaneous
rs10168	5:80654584	C	T	MSH3	1.40E-18	-0.41	Whole Blood
rs10168	5:80654584	C	T	MSH3	4.80E-16	-0.55	Cells - Transformed fibroblasts
rs10168	5:80654584	C	T	MSH3	1.80E-14	-0.4	Heart - Atrial Appendage
rs10168	5:80654584	C	T	MSH3	2.80E-14	-0.51	Artery - Aorta
rs10168	5:80654584	C	T	MSH3	8.20E-11	-0.33	Adipose - Visceral (Omentum)
rs10168	5:80654584	C	T	MSH3	1.50E-10	-0.44	Pancreas
rs10168	5:80654584	C	T	MSH3	3.70E-09	-0.36	Colon - Sigmoid
rs10168	5:80654584	C	T	MSH3	3.10E-08	-0.42	Adrenal Gland
rs10168	5:80654584	C	T	MSH3	7.50E-08	-0.35	Breast - Mammary Tissue
rs10168	5:80654584	C	T	MSH3	1.50E-07	-0.42	Pituitary
rs10168	5:80654584	C	T	MSH3	2.10E-07	-0.3	Colon - Transverse
rs10168	5:80654584	C	T	MSH3	5.70E-07	-0.34	Esophagus - Gastroesophageal Junction
rs10168	5:80654584	C	T	MSH3	6.80E-07	-0.26	Esophagus - Muscularis
rs10168	5:80654584	C	T	MSH3	7.90E-06	-0.18	Esophagus - Mucosa
rs2250063	5:80654678	C	T	DHFR	5.60E-92	-1	Muscle - Skeletal
rs2250063	5:80654678	C	T	DHFR	5.70E-51	-0.62	Adipose - Subcutaneous
rs2250063	5:80654678	C	T	DHFR	2.70E-48	-0.65	Whole Blood
rs2250063	5:80654678	C	T	DHFR	9.70E-46	-0.86	Testis
rs2250063	5:80654678	C	T	DHFR	4.10E-41	-0.57	Adipose - Visceral (Omentum)
rs2250063	5:80654678	C	T	DHFR	3.30E-33	-0.77	Heart - Left Ventricle
rs2250063	5:80654678	C	T	DHFR	1.80E-31	-0.3	Cells - Transformed fibroblasts
rs2250063	5:80654678	C	T	DHFR	1.30E-29	-0.82	Colon - Sigmoid
rs2250063	5:80654678	C	T	DHFR	4.90E-28	-0.69	Esophagus - Muscularis
rs2250063	5:80654678	C	T	DHFR	1.10E-26	-0.76	Heart - Atrial Appendage
rs2250063	5:80654678	C	T	DHFR	1.20E-22	-0.29	Esophagus - Mucosa
rs2250063	5:80654678	C	T	DHFR	1.40E-22	-0.51	Cells - EBV-transformed lymphocytes
rs2250063	5:80654678	C	T	DHFR	3.30E-20	-0.33	Lung
rs2250063	5:80654678	C	T	DHFR	5.30E-20	-0.42	Colon - Transverse
rs2250063	5:80654678	C	T	DHFR	1.90E-19	-0.95	Brain - Cortex
rs2250063	5:80654678	C	T	DHFR	3.80E-19	-0.42	Breast - Mammary Tissue
rs2250063	5:80654678	C	T	DHFR	1.00E-17	-0.92	Brain - Anterior cingulate cortex (BA24)
rs2250063	5:80654678	C	T	DHFR	4.00E-17	-0.32	Skin - Not Sun Exposed (Suprapubic)
rs2250063	5:80654678	C	T	DHFR	7.30E-16	-0.42	Artery - Tibial
rs2250063	5:80654678	C	T	DHFR	2.80E-15	-0.3	Skin - Sun Exposed (Lower leg)
rs2250063	5:80654678	C	T	DHFR	1.50E-14	-0.64	Brain - Nucleus accumbens (basal ganglia)
rs2250063	5:80654678	C	T	DHFR	3.20E-14	-0.43	Nerve - Tibial
rs2250063	5:80654678	C	T	DHFR	3.50E-14	-0.86	Brain - Frontal Cortex (BA9)
rs2250063	5:80654678	C	T	DHFR	1.00E-13	-0.66	Brain - Caudate (basal ganglia)
rs2250063	5:80654678	C	T	DHFR	1.90E-13	-0.52	Spleen
rs2250063	5:80654678	C	T	DHFR	2.10E-12	-0.56	Esophagus - Gastroesophageal Junction
rs2250063	5:80654678	C	T	DHFR	3.30E-12	-0.64	Brain - Hypothalamus
rs2250063	5:80654678	C	T	DHFR	3.80E-12	-0.75	Brain - Cerebellum
rs2250063	5:80654678	C	T	DHFR	3.80E-12	-0.63	Brain - Hippocampus
rs2250063	5:80654678	C	T	DHFR	1.20E-11	-0.75	Brain - Cerebellar Hemisphere
rs2250063	5:80654678	C	T	DHFR	2.80E-11	-0.62	Brain - Amygdala
rs2250063	5:80654678	C	T	DHFR	2.00E-10	-0.57	Adrenal Gland
rs2250063	5:80654678	C	T	DHFR	7.80E-09	-0.49	Pituitary
rs2250063	5:80654678	C	T	DHFR	1.50E-08	-0.73	Brain - Spinal cord (cervical c-1)
rs2250063	5:80654678	C	T	DHFR	2.60E-08	-0.6	Brain - Substantia nigra
rs2250063	5:80654678	C	T	DHFR	4.70E-08	-0.55	Prostate
rs2250063	5:80654678	C	T	DHFR	3.60E-07	-0.47	Artery - Coronary
rs2250063	5:80654678	C	T	DHFR	6.80E-07	-0.39	Artery - Aorta

rs2250063	5:80654678	C	T	MSH3	1.00E-27	-0.48	Thyroid
rs2250063	5:80654678	C	T	MSH3	9.90E-24	-0.41	Skin - Sun Exposed (Lower leg)
rs2250063	5:80654678	C	T	MSH3	1.40E-23	-0.49	Artery - Tibial
rs2250063	5:80654678	C	T	MSH3	7.20E-23	-0.45	Skin - Not Sun Exposed (Suprapubic)
rs2250063	5:80654678	C	T	MSH3	2.10E-19	-0.38	Adipose - Subcutaneous
rs2250063	5:80654678	C	T	MSH3	3.10E-19	-0.47	Nerve - Tibial
rs2250063	5:80654678	C	T	MSH3	9.60E-18	-0.4	Whole Blood
rs2250063	5:80654678	C	T	MSH3	2.40E-14	-0.52	Cells - Transformed fibroblasts
rs2250063	5:80654678	C	T	MSH3	1.40E-13	-0.4	Heart - Atrial Appendage
rs2250063	5:80654678	C	T	MSH3	3.10E-13	-0.5	Artery - Aorta
rs2250063	5:80654678	C	T	MSH3	4.70E-11	-0.33	Adipose - Visceral (Omentum)
rs2250063	5:80654678	C	T	MSH3	4.10E-10	-0.43	Pancreas
rs2250063	5:80654678	C	T	MSH3	3.50E-08	-0.35	Colon - Sigmoid
rs2250063	5:80654678	C	T	MSH3	3.70E-08	-0.41	Adrenal Gland
rs2250063	5:80654678	C	T	MSH3	1.90E-07	-0.42	Pituitary
rs2250063	5:80654678	C	T	MSH3	3.00E-07	-0.34	Breast - Mammary Tissue
rs2250063	5:80654678	C	T	MSH3	3.20E-07	-0.3	Colon - Transverse
rs2250063	5:80654678	C	T	MSH3	1.80E-06	-0.25	Esophagus - Muscularis
rs2250063	5:80654678	C	T	MSH3	2.70E-06	-0.32	Esophagus - Gastroesophageal Junction
rs2250063	5:80654678	C	T	MSH3	2.20E-05	-0.17	Esophagus - Mucosa
rs1105525	5:80654689	C	T	MSH3	4.00E-09	0.47	Cells - Transformed fibroblasts
rs1105525	5:80654689	C	T	MSH3	7.90E-07	0.26	Skin - Sun Exposed (Lower leg)
rs1105524	5:80654693	A	G	DHFR	1.80E-31	-0.66	Muscle - Skeletal
rs1105524	5:80654693	A	G	DHFR	1.20E-28	-0.58	Nerve - Tibial
rs1105524	5:80654693	A	G	DHFR	3.60E-28	-0.55	Artery - Tibial
rs1105524	5:80654693	A	G	DHFR	3.50E-27	-0.5	Adipose - Subcutaneous
rs1105524	5:80654693	A	G	DHFR	1.20E-24	-0.62	Esophagus - Muscularis
rs1105524	5:80654693	A	G	DHFR	1.20E-23	-0.7	Artery - Aorta
rs1105524	5:80654693	A	G	DHFR	2.80E-20	-0.67	Esophagus - Gastroesophageal Junction
rs1105524	5:80654693	A	G	DHFR	4.50E-19	-0.63	Heart - Atrial Appendage
rs1105524	5:80654693	A	G	DHFR	9.10E-19	-0.58	Heart - Left Ventricle
rs1105524	5:80654693	A	G	DHFR	2.80E-18	-0.29	Lung
rs1105524	5:80654693	A	G	DHFR	3.50E-15	-0.38	Breast - Mammary Tissue
rs1105524	5:80654693	A	G	DHFR	8.10E-15	-0.29	Skin - Sun Exposed (Lower leg)
rs1105524	5:80654693	A	G	DHFR	1.10E-14	-0.35	Adipose - Visceral (Omentum)
rs1105524	5:80654693	A	G	DHFR	6.10E-14	-0.42	Thyroid
rs1105524	5:80654693	A	G	DHFR	7.70E-14	-0.21	Esophagus - Mucosa
rs1105524	5:80654693	A	G	DHFR	3.50E-13	-0.2	Cells - Transformed fibroblasts
rs1105524	5:80654693	A	G	DHFR	1.70E-12	-0.66	Adrenal Gland
rs1105524	5:80654693	A	G	DHFR	5.90E-12	-0.33	Colon - Transverse
rs1105524	5:80654693	A	G	DHFR	1.20E-11	-0.35	Whole Blood
rs1105524	5:80654693	A	G	DHFR	2.90E-11	-0.79	Brain - Cerebellum
rs1105524	5:80654693	A	G	DHFR	1.50E-10	-0.54	Pituitary
rs1105524	5:80654693	A	G	DHFR	1.80E-10	-0.46	Spleen
rs1105524	5:80654693	A	G	DHFR	2.00E-10	-0.52	Colon - Sigmoid
rs1105524	5:80654693	A	G	DHFR	3.10E-10	-0.44	Testis
rs1105524	5:80654693	A	G	DHFR	1.70E-09	-0.76	Brain - Anterior cingulate cortex (BA24)
rs1105524	5:80654693	A	G	DHFR	4.40E-09	-0.68	Brain - Cerebellar Hemisphere
rs1105524	5:80654693	A	G	DHFR	8.30E-09	-0.22	Skin - Not Sun Exposed (Suprapubic)
rs1105524	5:80654693	A	G	DHFR	3.50E-08	-0.54	Brain - Caudate (basal ganglia)
rs1105524	5:80654693	A	G	DHFR	4.90E-08	-0.66	Brain - Frontal Cortex (BA9)
rs1105524	5:80654693	A	G	DHFR	2.50E-07	-0.53	Artery - Coronary
rs1105524	5:80654693	A	G	DHFR	3.20E-07	-0.6	Brain - Cortex
rs1105524	5:80654693	A	G	DHFR	1.00E-06	-0.54	Brain - Hypothalamus
rs1105524	5:80654693	A	G	DHFR	5.30E-06	-0.43	Brain - Putamen (basal ganglia)
rs1105524	5:80654693	A	G	MSH3	9.50E-11	0.38	Liver
rs1105524	5:80654693	A	G	MSH3	3.20E-07	0.17	Muscle - Skeletal
rs1650697	5:80654962	A	G	DHFR	1.50E-20	0.55	Thyroid
rs1650697	5:80654962	A	G	DHFR	8.20E-13	0.41	Artery - Tibial
rs1650697	5:80654962	A	G	DHFR	6.70E-12	0.55	Artery - Aorta
rs1650697	5:80654962	A	G	DHFR	3.50E-11	0.4	Nerve - Tibial
rs1650697	5:80654962	A	G	DHFR	4.60E-08	0.46	Esophagus - Gastroesophageal Junction
rs1650697	5:80654962	A	G	DHFR	5.70E-08	0.47	Pancreas
rs1650697	5:80654962	A	G	DHFR	9.90E-08	0.36	Esophagus - Muscularis
rs1650697	5:80654962	A	G	DHFR	2.20E-07	0.25	Stomach
rs1650697	5:80654962	A	G	DHFR	3.60E-06	0.2	Skin - Sun Exposed (Lower leg)
rs1650697	5:80654962	A	G	DHFR	9.70E-06	0.17	Lung
rs1650697	5:80654962	A	G	MSH3	2.80E-41	-0.83	Cells - Transformed fibroblasts
rs1650697	5:80654962	A	G	MSH3	1.20E-30	-0.51	Skin - Sun Exposed (Lower leg)
rs1650697	5:80654962	A	G	MSH3	3.80E-30	-0.54	Whole Blood
rs1650697	5:80654962	A	G	MSH3	6.30E-28	-0.53	Thyroid
rs1650697	5:80654962	A	G	MSH3	1.40E-25	-0.54	Artery - Tibial
rs1650697	5:80654962	A	G	MSH3	3.10E-25	-0.5	Skin - Not Sun Exposed (Suprapubic)
rs1650697	5:80654962	A	G	MSH3	7.00E-25	-0.56	Nerve - Tibial
rs1650697	5:80654962	A	G	MSH3	2.30E-24	-0.38	Esophagus - Mucosa
rs1650697	5:80654962	A	G	MSH3	8.30E-23	-0.42	Lung
rs1650697	5:80654962	A	G	MSH3	5.20E-22	-0.66	Artery - Aorta
rs1650697	5:80654962	A	G	MSH3	5.60E-19	-0.31	Muscle - Skeletal
rs1650697	5:80654962	A	G	MSH3	3.00E-18	-0.4	Adipose - Subcutaneous
rs1650697	5:80654962	A	G	MSH3	1.10E-17	-0.43	Adipose - Visceral (Omentum)
rs1650697	5:80654962	A	G	MSH3	1.40E-15	-0.29	Heart - Left Ventricle
rs1650697	5:80654962	A	G	MSH3	6.80E-15	-0.55	Colon - Sigmoid

rs1650697	5:80654962	A	G	MSH3	1.20E-14	-0.39	Esophagus - Muscularis
rs1650697	5:80654962	A	G	MSH3	1.90E-14	-0.46	Colon - Transverse
rs1650697	5:80654962	A	G	MSH3	6.70E-14	-0.6	Adrenal Gland
rs1650697	5:80654962	A	G	MSH3	7.60E-14	-0.5	Pancreas
rs1650697	5:80654962	A	G	MSH3	4.80E-13	-0.41	Heart - Atrial Appendage
rs1650697	5:80654962	A	G	MSH3	6.00E-13	-0.48	Esophagus - Gastroesophageal Junction
rs1650697	5:80654962	A	G	MSH3	7.20E-13	-0.53	Stomach
rs1650697	5:80654962	A	G	MSH3	3.60E-10	-0.57	Spleen
rs1650697	5:80654962	A	G	MSH3	4.30E-10	-0.41	Breast - Mammary Tissue
rs1650697	5:80654962	A	G	MSH3	4.20E-09	-0.36	Liver
rs1650697	5:80654962	A	G	MSH3	8.80E-08	-0.45	Artery - Coronary
rs1650697	5:80654962	A	G	MSH3	4.60E-07	-0.41	Prostate
rs1650697	5:80654962	A	G	MSH3	7.10E-06	-0.45	Small Intestine - Terminal Ileum
rs1650697	5:80654962	A	G	MSH3	1.50E-05	-0.41	Pituitary
rs1677658	5:80655040	G	T	DHFR	1.80E-26	-0.82	Muscle - Skeletal
rs1677658	5:80655040	G	T	DHFR	1.20E-17	-0.77	Testis
rs1677658	5:80655040	G	T	DHFR	1.60E-15	-0.56	Whole Blood
rs1677658	5:80655040	G	T	DHFR	1.40E-13	-0.75	Heart - Left Ventricle
rs1677658	5:80655040	G	T	DHFR	1.60E-12	-0.46	Adipose - Visceral (Omentum)
rs1677658	5:80655040	G	T	DHFR	2.70E-12	-0.28	Cells - Transformed fibroblasts
rs1677658	5:80655040	G	T	DHFR	2.00E-11	-0.73	Heart - Atrial Appendage
rs1677658	5:80655040	G	T	DHFR	3.10E-11	-0.3	Esophagus - Mucosa
rs1677658	5:80655040	G	T	DHFR	4.10E-10	-0.41	Adipose - Subcutaneous
rs1677658	5:80655040	G	T	DHFR	6.60E-09	-0.42	Artery - Tibial
rs1677658	5:80655040	G	T	DHFR	2.10E-08	-0.29	Lung
rs1677658	5:80655040	G	T	DHFR	6.60E-08	-0.35	Colon - Transverse
rs1677658	5:80655040	G	T	DHFR	1.20E-07	-0.51	Esophagus - Muscularis
rs1677658	5:80655040	G	T	DHFR	2.20E-07	-0.45	Cells - EBV-transformed lymphocytes
rs1677658	5:80655040	G	T	DHFR	3.90E-07	-0.41	Nerve - Tibial
rs1677658	5:80655040	G	T	DHFR	7.40E-07	-0.76	Brain - Cortex
rs1677658	5:80655040	G	T	DHFR	2.40E-06	-0.58	Colon - Sigmoid
rs1677658	5:80655040	G	T	DHFR	5.90E-06	-0.58	Brain - Caudate (basal ganglia)
rs1677658	5:80655040	G	T	DHFR	1.10E-05	-0.64	Brain - Cerebellar Hemisphere
rs1677658	5:80655040	G	T	DHFR	1.20E-05	-0.32	Breast - Mammary Tissue
rs1677658	5:80655040	G	T	DHFR	2.40E-05	-0.7	Brain - Cerebellum
rs1677658	5:80655040	G	T	MSH3	4.00E-15	-0.51	Thyroid
rs1677658	5:80655040	G	T	MSH3	4.20E-14	-0.51	Artery - Tibial
rs1677658	5:80655040	G	T	MSH3	4.50E-14	-0.5	Skin - Not Sun Exposed (Suprapubic)
rs1677658	5:80655040	G	T	MSH3	6.60E-14	-0.43	Skin - Sun Exposed (Lower leg)
rs1677658	5:80655040	G	T	MSH3	1.80E-13	-0.54	Nerve - Tibial
rs1677658	5:80655040	G	T	MSH3	6.30E-11	-0.4	Adipose - Subcutaneous
rs1677658	5:80655040	G	T	MSH3	1.60E-08	-0.56	Cells - Transformed fibroblasts
rs1677658	5:80655040	G	T	MSH3	5.40E-08	-0.37	Whole Blood
rs1677658	5:80655040	G	T	MSH3	1.90E-07	-0.36	Adipose - Visceral (Omentum)
rs1677658	5:80655040	G	T	MSH3	5.80E-07	-0.48	Breast - Mammary Tissue
rs1677658	5:80655040	G	T	MSH3	9.10E-07	-0.38	Heart - Atrial Appendage
rs1677658	5:80655040	G	T	MSH3	9.20E-07	-0.3	Lung
rs1677658	5:80655040	G	T	MSH3	1.40E-05	-0.42	Stomach
rs1677658	5:80655040	G	T	MSH3	2.60E-05	-0.25	Esophagus - Mucosa
rs1677658	5:80655040	G	T	MSH3	4.10E-05	-0.3	Esophagus - Muscularis

10.14 HD transcriptome-wide association study (TWAS)

Table 10.4. Transcriptome-wide association study (TWAS) of HD prefrontal cortex.

The method of Gusev et al. (2016) was used to test for association between phenotype and gene expression in control dorsolateral prefrontal cortex from the Common Mind Consortium (n=452) using summary statistics of genome-wide association studies. The Z-score represents the standardized effect size, with positive values indicating later onset (GeM-HD) or faster progression (TRACK-HD and REGISTRY). GeM – Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium GWAS (n = 4082) (GeM-HD, 2015). TRACK-HD and REGISTRY – Hensman Moss et al. (2017b) GWAS (n = 243). Table sorted by TRACK-HD, then TRACK-HD + REGISTRY, then GeM-HD p-values. MSH3 and FAN1 are highlighted red.

Gene symbol	TRACK-HD progression		TRACK-HD + REGISTRY progression		GeM-HD age at onset	
	Z	p	Z	p	Z	p
MSH3	4.71	2.52E-06	6.35	2.23E-10	-3.14	1.71E-03
DHFR	4.61	4.08E-06	3.51	4.44E-04	0.17	8.68E-01
THUMPD3	3.51	4.42E-04	3.35	8.01E-04	1.20	2.29E-01
UBE2G1	-3.51	4.55E-04	-	-	0.88	3.78E-01
RBM47	-3.47	5.30E-04	-3.19	1.41E-03	-0.74	4.60E-01
FBLL1	-3.43	5.99E-04	-	-	1.39	1.64E-01
WRB	-3.33	8.78E-04	-	-	-1.96	4.98E-02
TDH	-3.32	8.93E-04	-	-	-1.86	6.34E-02
ELOVL5	-3.24	1.22E-03	-	-	0.31	7.55E-01
RASGRF2	3.17	1.52E-03	-	-	-0.51	6.11E-01
PSMG1	3.15	1.64E-03	-	-	0.10	9.17E-01
LPL	-3.06	2.24E-03	-	-	-0.18	8.58E-01
ASPRV1	-3.05	2.25E-03	-	-	1.92	5.51E-02
C12orf43	3.05	2.29E-03	-	-	-0.53	5.95E-01
SLC15A5	3.05	2.30E-03	-	-	-0.85	3.96E-01
EFCAB5	3.01	2.63E-03	-	-	0.52	6.03E-01
KIAA1804	-3.00	2.67E-03	-3.37	7.44E-04	0.63	5.30E-01
TMEM132B	2.99	2.82E-03	-	-	0.18	8.60E-01
SLC25A37	-2.97	2.94E-03	-	-	0.31	7.59E-01
FAHD1	-2.97	2.96E-03	-	-	-0.59	5.56E-01
CD93	2.94	3.30E-03	-	-	-1.19	2.34E-01
RNF126	2.92	3.54E-03	-	-	0.47	6.42E-01
ORC5	-2.90	3.78E-03	-	-	0.42	6.75E-01
FBRSL1	2.88	3.97E-03	-	-	0.17	8.66E-01
LOC283683	2.86	4.20E-03	-	-	-1.28	2.01E-01
INTS12	-2.86	4.22E-03	-	-	0.89	3.75E-01
ADAMTS19	2.86	4.23E-03	-	-	-1.56	1.19E-01
RNASE4	2.85	4.32E-03	-	-	-0.60	5.47E-01
GPR124	2.85	4.35E-03	2.90	3.78E-03	-1.45	1.46E-01
RNF24	2.82	4.85E-03	-	-	-0.11	9.14E-01
DNTTIP1	2.82	4.87E-03	-	-	-1.67	9.43E-02
NRG4	2.79	5.25E-03	-	-	-0.43	6.64E-01
KLHDC4	2.78	5.43E-03	-	-	0.02	9.83E-01
NDUFS5	-2.77	5.68E-03	-	-	-0.24	8.06E-01
UBE2G2	-2.76	5.74E-03	-	-	-1.16	2.47E-01
GALNT3	2.75	5.97E-03	-	-	-0.82	4.13E-01
OMA1	-2.73	6.29E-03	-	-	0.01	9.91E-01
SSH2	2.73	6.30E-03	-	-	0.72	4.74E-01
ATP1B3	2.72	6.59E-03	3.05	2.32E-03	-0.35	7.29E-01
NCEH1	-2.70	6.86E-03	-	-	-1.05	2.92E-01
OGN	-2.70	6.94E-03	-	-	0.04	9.65E-01
PHLDA3	2.70	7.03E-03	-	-	0.48	6.30E-01
GSTM3	2.69	7.05E-03	-	-	-1.03	3.02E-01
CRYZ	-2.68	7.28E-03	-	-	-0.37	7.14E-01
RBM23	2.68	7.39E-03	-	-	0.72	4.75E-01
DNAJB11	2.68	7.47E-03	-	-	1.00	3.18E-01
CBLN1	-2.67	7.59E-03	-	-	-1.59	1.12E-01
HEATR1	2.66	7.83E-03	-	-	0.67	5.05E-01
CACNA1B	2.65	7.97E-03	-	-	-1.95	5.15E-02
PTRF	2.63	8.59E-03	-	-	-1.53	1.26E-01
PKNOX1	2.63	8.61E-03	-	-	0.03	9.73E-01
ACSL5	-2.62	8.72E-03	-	-	-1.83	6.76E-02
CCDC53	2.62	8.77E-03	3.20	1.37E-03	-0.99	3.20E-01
LOC339803	2.61	9.10E-03	-	-	-0.33	7.44E-01

ARHGAP22	2.60	9.34E-03	-	-	-1.97	4.94E-02
CNKSR3	-2.60	9.40E-03	-	-	-0.26	7.94E-01
LRRFIP1	-2.60	9.42E-03	-	-	-0.14	8.91E-01
PINX1	2.59	9.69E-03	-	-	-0.59	5.56E-01
PLSCR1	-2.58	9.77E-03	-	-	0.99	3.22E-01
PCDHGB4	-2.58	9.80E-03	-	-	1.15	2.49E-01
IFT57	-2.58	9.81E-03	-	-	-0.76	4.45E-01
FAM120B	-2.58	9.90E-03	-	-	-3.23	1.22E-03
POLD2	-2.57	1.02E-02	-	-	-0.72	4.70E-01
KLC3	2.55	1.07E-02	-	-	0.10	9.21E-01
EIF2B1	-2.55	1.09E-02	-	-	0.13	9.01E-01
MRGPRF	-2.54	1.09E-02	-	-	-0.55	5.82E-01
ZNF131	2.54	1.12E-02	-	-	0.88	3.81E-01
PIGN	-2.53	1.14E-02	-	-	-0.71	4.79E-01
TMEM41B	-2.53	1.14E-02	-	-	-0.33	7.43E-01
ARL3	-2.52	1.18E-02	-	-	-1.02	3.06E-01
INTS10	-2.52	1.18E-02	-	-	-1.68	9.31E-02
CYP27C1	-2.51	1.19E-02	-	-	0.79	4.27E-01
PKIB	-2.51	1.20E-02	-	-	-0.36	7.22E-01
SNRNP27	-2.51	1.21E-02	-	-	-1.29	1.97E-01
UBTD2	2.50	1.24E-02	-	-	-0.56	5.79E-01
UBE2Q2	2.49	1.28E-02	-	-	-0.69	4.90E-01
KRBA1	-2.49	1.28E-02	-	-	1.08	2.81E-01
HKDC1	2.48	1.32E-02	-	-	0.34	7.34E-01
CCND2	-2.47	1.34E-02	-	-	1.47	1.42E-01
CTSC	2.47	1.36E-02	-	-	-1.54	1.24E-01
LGALS8	-2.47	1.37E-02	-	-	-1.18	2.38E-01
PCDHGA6	2.46	1.39E-02	-	-	-1.21	2.27E-01
REEP3	-2.46	1.40E-02	-	-	0.52	6.03E-01
SDHA	2.45	1.41E-02	-	-	0.18	8.55E-01
BMPR1A	2.45	1.41E-02	-	-	0.99	3.24E-01
CLEC4GP1	2.45	1.42E-02	-	-	0.37	7.14E-01
ZNF770	-2.45	1.43E-02	-	-	0.93	3.54E-01
TRIM44	2.44	1.45E-02	-	-	-0.38	7.03E-01
PACRGL	2.44	1.45E-02	-	-	-1.70	9.00E-02
CBX7	-2.44	1.46E-02	-	-	0.09	9.31E-01
VEZT	2.44	1.46E-02	-	-	-0.08	9.33E-01
FAM124B	2.44	1.47E-02	-	-	-1.57	1.17E-01
C1orf228	2.44	1.47E-02	-	-	1.75	7.94E-02
HAS1	2.44	1.48E-02	-	-	-0.42	6.78E-01
ALX4	-2.43	1.49E-02	-2.94	3.26E-03	1.04	2.99E-01
ARHGEF11	-2.42	1.54E-02	-	-	0.12	9.02E-01
CSTF2T	2.42	1.54E-02	-	-	3.22	1.26E-03
DENND1B	2.42	1.55E-02	-	-	1.16	2.47E-01
LOC284751	2.41	1.59E-02	-	-	0.18	8.58E-01
PDGFD	-2.40	1.62E-02	-	-	-0.01	9.92E-01
RAB37	-2.40	1.64E-02	-	-	-0.22	8.22E-01
PANK2	2.39	1.68E-02	-	-	0.53	5.98E-01
LOC401321	2.39	1.70E-02	-	-	-0.84	3.99E-01
NUP107	-2.38	1.73E-02	-	-	-1.21	2.27E-01
IGFALS	2.37	1.76E-02	-	-	0.37	7.11E-01
PNPLA5	2.37	1.77E-02	-	-	0.31	7.56E-01
LMOD1	-2.37	1.78E-02	-	-	-0.45	6.51E-01
CCDC127	2.37	1.79E-02	-	-	0.26	7.92E-01
CYB5D2	2.37	1.80E-02	-	-	-1.18	2.39E-01
CERS2	-2.36	1.80E-02	-	-	1.93	5.40E-02
PLCB3	-2.36	1.81E-02	-	-	-0.98	3.26E-01
SMYD2	2.36	1.82E-02	-	-	-1.65	9.99E-02
ZNF667	-2.36	1.84E-02	-3.41	6.58E-04	-1.60	1.09E-01
ZNHIT6	2.36	1.84E-02	-	-	1.67	9.58E-02
MYO16	-2.36	1.85E-02	-	-	-1.21	2.28E-01
ZMYM5	2.35	1.85E-02	-	-	0.90	3.69E-01
LIN37	2.36	1.85E-02	-	-	0.56	5.75E-01
NMD3	-2.35	1.85E-02	-	-	-0.16	8.70E-01
FN3KRP	-2.35	1.87E-02	-	-	0.78	4.35E-01
ADPRH	-2.35	1.87E-02	-	-	-0.25	8.02E-01
ACTR1A	-2.35	1.89E-02	-	-	0.30	7.62E-01

ECE2	2.34	1.91E-02	-	-	0.42	6.72E-01
YPEL1	2.34	1.93E-02	-	-	-0.96	3.38E-01
IDI2-AS1	2.33	1.97E-02	-	-	-0.31	7.58E-01
RILPL1	-2.33	1.98E-02	-	-	0.59	5.56E-01
LOC100505761	2.33	1.98E-02	-	-	1.01	3.12E-01
BCL2L2	-2.33	2.00E-02	-	-	-1.70	8.97E-02
HMGN1	2.33	2.01E-02	-	-	0.82	4.12E-01
RRP1B	-2.32	2.03E-02	-	-	0.94	3.45E-01
KLHDC1	2.32	2.04E-02	-	-	-0.22	8.28E-01
TMEM63A	2.32	2.04E-02	-	-	0.03	9.79E-01
SNRPC	-2.32	2.05E-02	-	-	-1.27	2.05E-01
LPIN3	2.31	2.06E-02	-	-	-0.97	3.31E-01
RSBN1	2.31	2.08E-02	-	-	-0.16	8.74E-01
UHRF1BP1	2.31	2.08E-02	-	-	1.36	1.74E-01
TTC38	-2.31	2.09E-02	-	-	-0.12	9.02E-01
LCTL	-2.31	2.11E-02	-	-	1.04	2.98E-01
TAF9	2.30	2.14E-02	-	-	-0.20	8.43E-01
DSC1	-2.30	2.14E-02	-	-	0.16	8.73E-01
HAP1	-2.30	2.14E-02	-	-	-0.12	9.02E-01
HERC2	2.30	2.16E-02	-	-	0.85	3.95E-01
ADARB1	2.29	2.18E-02	-	-	0.17	8.63E-01
PDIA4	2.29	2.19E-02	-	-	0.42	6.74E-01
RAB6C	2.29	2.19E-02	-	-	-0.98	3.27E-01
MIA3	2.29	2.19E-02	-	-	0.34	7.30E-01
CPS1	2.29	2.21E-02	-	-	0.40	6.89E-01
RBL2	2.29	2.23E-02	-	-	0.09	9.25E-01
CCDC110	-2.28	2.24E-02	-	-	0.76	4.48E-01
MRPS21	-2.28	2.24E-02	-	-	0.13	8.96E-01
SPSB4	2.28	2.25E-02	3.34	8.37E-04	-1.10	2.71E-01
TTC30A	2.28	2.27E-02	-	-	0.79	4.32E-01
TMED10	2.28	2.28E-02	-	-	-2.21	2.68E-02
PEX11A	-2.28	2.28E-02	-	-	0.29	7.71E-01
ZNF497	-2.27	2.31E-02	-	-	-0.07	9.44E-01
CUBN	-2.27	2.34E-02	-	-	-0.89	3.71E-01
MRPL53	-2.26	2.36E-02	-	-	-1.41	1.59E-01
PKDREJ	2.26	2.36E-02	-	-	-0.44	6.59E-01
SLC17A6	2.26	2.36E-02	-	-	0.03	9.74E-01
SNHG8	-2.26	2.37E-02	-	-	1.66	9.70E-02
EIF2AK1	-2.26	2.41E-02	-	-	3.38	7.31E-04
CCDC146	2.25	2.47E-02	-	-	-1.29	1.97E-01
MAPK13	-2.25	2.47E-02	-	-	0.99	3.20E-01
ZNF514	2.24	2.49E-02	-	-	-0.02	9.80E-01
RNH1	2.24	2.50E-02	-	-	-1.26	2.07E-01
SPON1	-2.24	2.50E-02	-	-	0.11	9.15E-01
TMCO3	-2.24	2.51E-02	-	-	-1.71	8.70E-02
DSE	-2.24	2.52E-02	-	-	-0.89	3.72E-01
CFTR	2.24	2.54E-02	-	-	-0.40	6.91E-01
CNKSR1	2.23	2.55E-02	-	-	0.37	7.15E-01
HTR1A	-2.22	2.63E-02	-	-	-1.11	2.67E-01
FLJ20021	-2.22	2.65E-02	-	-	-0.30	7.62E-01
SLC16A10	-2.21	2.69E-02	-2.92	3.46E-03	0.12	9.02E-01
SNF8	2.21	2.69E-02	-	-	-2.50	1.24E-02
RTBDN	2.21	2.69E-02	3.31	9.29E-04	-0.81	4.18E-01
HPSE2	-2.21	2.71E-02	-	-	-0.53	5.95E-01
CKMT1B	2.21	2.73E-02	-	-	0.40	6.89E-01
CENPP	-2.21	2.73E-02	-	-	0.41	6.85E-01
CCDC82	-2.21	2.73E-02	-3.33	8.75E-04	2.32	2.05E-02
MCTP1	-2.21	2.73E-02	-	-	-0.27	7.90E-01
CRIPAK	-2.21	2.75E-02	-	-	-0.05	9.64E-01
RSPO2	-2.20	2.75E-02	-	-	0.18	8.60E-01
GNPTAB	-2.20	2.76E-02	-	-	-0.35	7.25E-01
KRBA2	-2.20	2.78E-02	-	-	-0.82	4.11E-01
KIF24	2.20	2.80E-02	-	-	-1.30	1.94E-01
CCDC102B	2.20	2.80E-02	-	-	0.22	8.26E-01
GEMIN4	2.19	2.86E-02	-	-	-0.11	9.15E-01
LOC100128288	-2.19	2.86E-02	-	-	-0.72	4.73E-01
VRK1	2.19	2.87E-02	-	-	-1.27	2.06E-01

KIF17	2.19	2.87E-02	-	-	0.63	5.27E-01
CES1	-2.19	2.89E-02	-	-	1.58	1.14E-01
GABRA1	2.18	2.90E-02	-	-	-0.38	7.08E-01
IGFLR1	-2.18	2.94E-02	-	-	-0.60	5.51E-01
BCL2L13	-2.18	2.94E-02	-	-	-2.26	2.37E-02
BTBD8	-2.18	2.94E-02	-	-	1.42	1.55E-01
ADSSL1	2.17	2.98E-02	-	-	-1.17	2.41E-01
C6orf165	2.17	3.00E-02	-	-	-0.31	7.59E-01
ABCC11	-2.17	3.01E-02	-	-	1.30	1.94E-01
VASH2	-2.17	3.03E-02	-	-	-0.31	7.55E-01
RTP4	-2.16	3.04E-02	-	-	1.01	3.14E-01
GUSBP11	2.16	3.06E-02	-	-	1.28	2.01E-01
TSNARE1	2.16	3.10E-02	-	-	-2.32	2.05E-02
YWHAH	-2.16	3.10E-02	-	-	0.00	9.98E-01
PELI2	2.16	3.10E-02	-	-	-1.37	1.71E-01
TMTC3	2.15	3.12E-02	-	-	1.31	1.90E-01
C6orf62	-2.15	3.13E-02	-	-	-0.72	4.73E-01
ANKRD40	2.14	3.26E-02	-	-	1.20	2.30E-01
FAM201A	-2.13	3.31E-02	-	-	1.20	2.29E-01
FAM83H	-2.13	3.32E-02	-	-	-2.11	3.49E-02
ZBTB41	2.13	3.32E-02	-	-	0.64	5.22E-01
ADCY3	2.13	3.33E-02	-	-	0.21	8.32E-01
NEFL	-2.13	3.34E-02	-	-	-1.40	1.62E-01
NBAS	-2.13	3.34E-02	-	-	-0.83	4.07E-01
TRIM9	-2.13	3.34E-02	-	-	0.75	4.53E-01
CISD2	2.13	3.35E-02	-	-	-0.05	9.59E-01
PLCD1	2.12	3.38E-02	-	-	0.99	3.22E-01
EEF2K	-2.12	3.40E-02	-	-	-1.16	2.45E-01
EIF2B2	-2.12	3.44E-02	-	-	2.62	8.81E-03
ZNF132	-2.11	3.46E-02	-	-	-0.04	9.70E-01
URB1	-2.11	3.46E-02	-	-	0.02	9.87E-01
VSTM4	-2.11	3.47E-02	-	-	-0.62	5.34E-01
BNIP1	-2.11	3.48E-02	-	-	-1.67	9.57E-02
MYCBPAP	-2.11	3.50E-02	-	-	-0.92	3.56E-01
SGCE	2.11	3.51E-02	-	-	-1.09	2.77E-01
C10orf67	2.10	3.55E-02	-	-	0.95	3.42E-01
MUL1	2.10	3.58E-02	-	-	0.14	8.86E-01
CCNG1	-2.10	3.58E-02	-	-	0.60	5.48E-01
HTR1F	2.09	3.64E-02	-	-	-0.20	8.43E-01
OAF	-2.09	3.65E-02	-	-	-0.36	7.16E-01
CYFIP1	2.09	3.66E-02	-	-	1.53	1.25E-01
LOC100131320	2.09	3.67E-02	-	-	-0.94	3.48E-01
SETD8	2.09	3.67E-02	-	-	-0.70	4.87E-01
RPS5	2.09	3.69E-02	-	-	-1.00	3.19E-01
REEP6	2.09	3.69E-02	-	-	-1.12	2.64E-01
SELRC1	2.08	3.74E-02	-	-	0.16	8.76E-01
KIAA1586	2.08	3.76E-02	-	-	-0.17	8.63E-01
SLC33A1	-2.08	3.76E-02	-	-	-1.02	3.08E-01
EIF3E	2.08	3.78E-02	-	-	0.70	4.87E-01
LOC100128361	-2.08	3.79E-02	-	-	0.47	6.38E-01
STX12	-2.07	3.84E-02	-	-	-0.38	7.06E-01
SGCB	2.07	3.86E-02	-	-	1.67	9.55E-02
DLG5	-2.07	3.87E-02	-	-	0.80	4.23E-01
ZNF519	2.07	3.88E-02	-	-	-1.49	1.37E-01
MPZL1	2.07	3.88E-02	-	-	0.62	5.39E-01
SESTD1	2.07	3.89E-02	-	-	-1.42	1.54E-01
FAM129B	2.07	3.89E-02	-	-	-1.44	1.50E-01
NDUFA2	2.06	3.93E-02	-	-	-1.16	2.45E-01
COL6A1	-2.06	3.94E-02	-	-	-0.11	9.15E-01
IFI16	-2.06	3.98E-02	-	-	0.11	9.10E-01
PGAP3	-2.05	4.00E-02	-	-	0.82	4.11E-01
PLEKHA1	-2.05	4.03E-02	-	-	-1.00	3.16E-01
ZNF445	-2.05	4.03E-02	-	-	-1.28	2.02E-01
RASA4	-2.05	4.05E-02	-	-	-1.38	1.67E-01
CHKB	-2.05	4.09E-02	-	-	-0.79	4.30E-01
HN1L	2.04	4.09E-02	-	-	-0.39	6.94E-01
CYP1B1-AS1	2.04	4.10E-02	-	-	-0.66	5.10E-01

GAS5	-2.04	4.14E-02	-	-	-2.03	4.23E-02
KIAA0825	-2.04	4.14E-02	-	-	-0.50	6.18E-01
SCARB2	2.04	4.16E-02	-	-	-0.05	9.63E-01
TXK	-2.04	4.16E-02	-	-	-0.93	3.54E-01
FLVCR1	2.03	4.23E-02	2.94	3.25E-03	-0.42	6.72E-01
VAMP1	2.03	4.23E-02	-	-	0.25	8.06E-01
AKTIP	2.03	4.24E-02	-	-	0.22	8.24E-01
GOSR1	-2.03	4.25E-02	-	-	-1.04	2.98E-01
SLC26A1	2.03	4.25E-02	-	-	-0.04	9.69E-01
APRT	-2.03	4.25E-02	-	-	-0.15	8.83E-01
SLC22A3	-2.03	4.26E-02	-	-	-1.38	1.66E-01
DENND1A	-2.02	4.31E-02	-	-	-0.53	5.99E-01
VILL	-2.02	4.32E-02	-	-	-2.06	3.98E-02
SOD2	-2.02	4.33E-02	-	-	-0.19	8.46E-01
PTPN3	-2.02	4.34E-02	-	-	0.38	7.07E-01
PSMB5	-2.02	4.34E-02	-	-	-1.17	2.40E-01
CD52	2.02	4.34E-02	-	-	-0.46	6.44E-01
ABCA11P	2.01	4.46E-02	-	-	-0.71	4.78E-01
JAK2	2.01	4.48E-02	-	-	-0.09	9.25E-01
WDR76	-2.01	4.49E-02	-	-	0.23	8.22E-01
MED13	-2.00	4.51E-02	-	-	0.55	5.82E-01
BTBD10	2.00	4.57E-02	-	-	-0.66	5.10E-01
C10orf32	2.00	4.57E-02	-	-	0.25	8.03E-01
YES1	2.00	4.57E-02	-	-	0.12	9.06E-01
CENPQ	-2.00	4.58E-02	-	-	0.62	5.33E-01
SFRP5	-2.00	4.60E-02	-	-	0.67	5.03E-01
MBLAC1	-1.99	4.61E-02	-	-	-1.64	1.02E-01
GPRC5B	1.99	4.61E-02	-	-	-1.57	1.16E-01
LOC441242	-1.99	4.64E-02	-	-	0.82	4.10E-01
PRKCB	1.99	4.69E-02	-	-	0.73	4.63E-01
SLC36A4	-1.99	4.70E-02	-	-	0.49	6.26E-01
ZNF593	-1.99	4.71E-02	-	-	-0.14	8.88E-01
ADAM17	1.98	4.72E-02	-	-	0.29	7.74E-01
MLH3	-1.98	4.73E-02	-	-	2.64	8.36E-03
SLC26A11	1.98	4.75E-02	-	-	0.81	4.19E-01
MPP6	1.98	4.76E-02	-	-	-0.67	5.03E-01
DENND3	1.98	4.78E-02	-	-	-0.36	7.19E-01
SYNE2	1.98	4.79E-02	-	-	-0.41	6.79E-01
HELB	-1.98	4.80E-02	-	-	-0.80	4.26E-01
MOV10L1	1.98	4.80E-02	-	-	-0.43	6.65E-01
BRD3	1.98	4.81E-02	-	-	-0.69	4.88E-01
PKDCC	-1.98	4.81E-02	-	-	0.25	8.03E-01
UBE2Q2P1	1.98	4.81E-02	-	-	-0.13	8.98E-01
ZNF584	1.98	4.81E-02	-	-	-0.59	5.56E-01
B4GALT4	1.97	4.85E-02	-	-	-0.74	4.60E-01
NUBP2	1.97	4.85E-02	-	-	0.25	8.03E-01
FDXR	1.97	4.86E-02	-	-	0.37	7.12E-01
SGK223	-1.97	4.88E-02	-	-	0.76	4.48E-01
C21orf128	1.97	4.89E-02	-	-	1.40	1.63E-01
RPF1	-1.97	4.90E-02	-	-	1.91	5.58E-02
PSD4	-1.97	4.92E-02	-	-	0.63	5.31E-01
FAM179B	1.97	4.92E-02	-	-	1.07	2.84E-01
SEMA4F	1.96	4.96E-02	-	-	0.49	6.24E-01
INPP1	1.96	4.98E-02	-	-	-0.71	4.80E-01
ABCC3	-1.96	4.99E-02	-	-	-1.05	2.93E-01
C1orf52	-1.96	5.00E-02	-	-	-2.61	8.93E-03
FAN1	-	-	-3.78	1.58E-04	6.99	2.80E-12
CUTC	-	-	-3.12	1.81E-03	1.09	2.77E-01
LRP2BP	-	-	-3.05	2.32E-03	-0.43	6.68E-01
ZNF492	-	-	3.02	2.53E-03	-0.46	6.45E-01
EPM2AIP1	-	-	3.00	2.66E-03	-0.68	4.97E-01
UFSP2	-	-	-3.00	2.73E-03	-1.61	1.06E-01
LEFTY1	-	-	2.94	3.27E-03	-0.53	5.95E-01
SGCD	-	-	-	-	4.30	1.70E-05
SUMF2	-	-	-	-	3.92	8.99E-05
GPR161	-	-	-	-	3.55	3.86E-04
ARID3B	-	-	-	-	3.35	8.19E-04

DDX20	-	-	-	-	3.22	1.28E-03
PRKG1	-	-	-	-	3.22	1.28E-03
PMS2	-	-	-	-	3.20	1.39E-03
RBM6	-	-	-	-	3.17	1.53E-03
CCZ1	-	-	-	-	-3.15	1.62E-03
PMS1	-	-	-	-	3.10	1.91E-03
NEU3	-	-	-	-	3.08	2.08E-03
PRICKLE1	-	-	-	-	-3.08	2.10E-03
B3GAT1	-	-	-	-	-3.07	2.11E-03
TRANK1	-	-	-	-	-3.07	2.15E-03
MST1R	-	-	-	-	-3.02	2.51E-03
NEDD4	-	-	-	-	-2.98	2.88E-03
SNCAIP	-	-	-	-	2.96	3.03E-03
DPM2	-	-	-	-	-2.96	3.06E-03
FRRS1	-	-	-	-	-2.94	3.24E-03
GMPPB	-	-	-	-	-2.94	3.26E-03
LYPLAL1	-	-	-	-	-2.91	3.58E-03
OXGR1	-	-	-	-	2.91	3.67E-03
SQSTM1	-	-	-	-	2.90	3.77E-03
DGKE	-	-	-	-	2.87	4.14E-03
SNAP91	-	-	-	-	2.86	4.20E-03
SLC4A8	-	-	-	-	-2.86	4.21E-03
SLCO2B1	-	-	-	-	-2.84	4.48E-03
SURF4	-	-	-	-	2.84	4.54E-03
NHLRC1	-	-	-	-	-2.83	4.61E-03
DBT	-	-	-	-	2.82	4.86E-03
PSMA4	-	-	-	-	2.82	4.88E-03
TECPR2	-	-	-	-	2.78	5.39E-03
ADAMTS3	-	-	-	-	2.77	5.59E-03
MCAT	-	-	-	-	-2.77	5.62E-03
IFITM10	-	-	-	-	2.77	5.64E-03
PTPRK	-	-	-	-	2.76	5.70E-03
BCL2L11	-	-	-	-	-2.73	6.31E-03
SULT1A1	-	-	-	-	2.73	6.36E-03
PLEKHG5	-	-	-	-	-2.71	6.65E-03
C5orf45	-	-	-	-	2.71	6.83E-03
GOLIM4	-	-	-	-	-2.70	7.03E-03
DCLK3	-	-	-	-	-2.69	7.20E-03
LCA5L	-	-	-	-	-2.68	7.30E-03
FMO4	-	-	-	-	2.68	7.32E-03
PRTFDC1	-	-	-	-	2.68	7.40E-03
CREG2	-	-	-	-	2.58	9.84E-03
TMC5	-	-	-	-	2.58	9.95E-03
GALNT2	-	-	-	-	-2.57	1.02E-02
LRRFIP2	-	-	-	-	2.56	1.04E-02
VOPP1	-	-	-	-	2.56	1.06E-02
TRAF4	-	-	-	-	-2.55	1.08E-02
ZNF280D	-	-	-	-	-2.54	1.11E-02
RHPN1	-	-	-	-	-2.53	1.13E-02
HPS4	-	-	-	-	-2.53	1.14E-02
SNUPN	-	-	-	-	2.53	1.14E-02
TTC21A	-	-	-	-	2.52	1.18E-02
SNHG1	-	-	-	-	-2.51	1.19E-02
ZNF329	-	-	-	-	2.52	1.19E-02
C6orf120	-	-	-	-	-2.49	1.27E-02
SLCO3A1	-	-	-	-	2.49	1.29E-02
KIAA0040	-	-	-	-	-2.48	1.30E-02
LOC100131564	-	-	-	-	2.48	1.31E-02
C1orf229	-	-	-	-	2.48	1.32E-02
USP46	-	-	-	-	-2.48	1.32E-02
CHRNA5	-	-	-	-	2.47	1.35E-02
PCOLCE2	-	-	-	-	-2.46	1.39E-02
SCRG1	-	-	-	-	-2.45	1.42E-02
NUPL2	-	-	-	-	2.45	1.43E-02
ABHD14B	-	-	-	-	2.44	1.46E-02
C10orf18	-	-	-	-	-2.44	1.46E-02
NWD1	-	-	-	-	-2.44	1.46E-02

ACBD7	-	-	-	-	-2.44	1.47E-02
LOC283177	-	-	-	-	-2.44	1.47E-02
POLR2J4	-	-	-	-	-2.43	1.51E-02
PIGX	-	-	-	-	-2.43	1.52E-02
PRKD1	-	-	-	-	-2.42	1.55E-02
S1PR2	-	-	-	-	2.42	1.57E-02
CCDC102A	-	-	-	-	2.41	1.58E-02
RPA3	-	-	-	-	-2.41	1.59E-02
PPP1CB	-	-	-	-	2.41	1.61E-02
COMMD4	-	-	-	-	-2.40	1.63E-02
MEF2BNB	-	-	-	-	-2.40	1.64E-02
ZNF75A	-	-	-	-	-2.38	1.72E-02
BCR	-	-	-	-	-2.38	1.73E-02
CCT6P1	-	-	-	-	2.38	1.73E-02
TXNDC12	-	-	-	-	2.38	1.75E-02
ORMDL1	-	-	-	-	-2.37	1.76E-02
STAG3L4	-	-	-	-	2.37	1.80E-02
GPX1	-	-	-	-	-2.36	1.81E-02
MCM7	-	-	-	-	2.36	1.82E-02
MRPL24	-	-	-	-	2.34	1.91E-02
HEATR2	-	-	-	-	-2.34	1.93E-02
RHOA	-	-	-	-	-2.34	1.93E-02
CYP24A1	-	-	-	-	-2.34	1.94E-02
ASNSD1	-	-	-	-	2.34	1.95E-02
FIG4	-	-	-	-	-2.33	1.96E-02
CBS	-	-	-	-	2.33	1.97E-02
LRPPRC	-	-	-	-	-2.33	1.99E-02
ENO3	-	-	-	-	2.32	2.01E-02
QRICH2	-	-	-	-	2.32	2.03E-02
KIAA1731	-	-	-	-	2.32	2.04E-02
GPR179	-	-	-	-	2.32	2.05E-02
WDR36	-	-	-	-	-2.31	2.06E-02
ANO7	-	-	-	-	-2.31	2.10E-02
C6orf70	-	-	-	-	-2.31	2.10E-02
LYPD6	-	-	-	-	-2.31	2.10E-02
TRIOBP	-	-	-	-	-2.30	2.13E-02
ECM2	-	-	-	-	2.29	2.21E-02
IFT52	-	-	-	-	-2.29	2.21E-02
RRP7B	-	-	-	-	-2.28	2.25E-02
LEMD3	-	-	-	-	2.28	2.26E-02
GNPDA2	-	-	-	-	2.28	2.28E-02
RIIAD1	-	-	-	-	2.28	2.28E-02
INCA1	-	-	-	-	2.28	2.29E-02
SLCO4C1	-	-	-	-	2.28	2.29E-02
GAN	-	-	-	-	-2.27	2.31E-02
LRRC1	-	-	-	-	2.27	2.31E-02
TMEM165	-	-	-	-	-2.27	2.34E-02
ZNF334	-	-	-	-	-2.26	2.38E-02
MAP3K4	-	-	-	-	-2.26	2.40E-02
RPP38	-	-	-	-	2.25	2.42E-02
STEAP1	-	-	-	-	-2.25	2.43E-02
LOC285889	-	-	-	-	2.25	2.44E-02
RARRES1	-	-	-	-	-2.25	2.44E-02
DGCR2	-	-	-	-	-2.25	2.45E-02
ZKSCAN1	-	-	-	-	2.25	2.45E-02
ZNF669	-	-	-	-	-2.24	2.50E-02
PARD3	-	-	-	-	2.24	2.51E-02
ZNF639	-	-	-	-	2.24	2.52E-02
ADCK3	-	-	-	-	2.23	2.60E-02
NANS	-	-	-	-	-2.23	2.60E-02
PPAT	-	-	-	-	-2.23	2.60E-02
ABCC6P1	-	-	-	-	-2.22	2.64E-02
MDGA1	-	-	-	-	2.21	2.71E-02
ZNF681	-	-	-	-	-2.21	2.71E-02
GLOD4	-	-	-	-	-2.21	2.73E-02
ZFAT	-	-	-	-	-2.20	2.77E-02
ZNF713	-	-	-	-	-2.20	2.77E-02

METTL21D	-	-	-	-	-2.20	2.78E-02
ZFP1	-	-	-	-	2.20	2.78E-02
TMCC3	-	-	-	-	2.19	2.82E-02
MCM8	-	-	-	-	2.19	2.84E-02
NFATC2IP	-	-	-	-	-2.19	2.84E-02
PCP4L1	-	-	-	-	-2.19	2.84E-02
IPO11	-	-	-	-	-2.19	2.86E-02
FRG1	-	-	-	-	-2.18	2.89E-02
PRKD3	-	-	-	-	2.18	2.90E-02
P4HTM	-	-	-	-	2.18	2.91E-02
ZNF814	-	-	-	-	-2.18	2.94E-02
RYK	-	-	-	-	-2.18	2.96E-02
TDRD9	-	-	-	-	2.17	2.97E-02
ZNF670	-	-	-	-	-2.17	2.97E-02
PAQR4	-	-	-	-	2.17	3.03E-02
NCBP2	-	-	-	-	2.16	3.05E-02
ERCC3	-	-	-	-	-2.16	3.08E-02
GPR133	-	-	-	-	2.16	3.10E-02
OR2D3	-	-	-	-	-2.16	3.10E-02
SLC27A2	-	-	-	-	-2.16	3.11E-02
C17orf107	-	-	-	-	2.15	3.12E-02
PSMB6	-	-	-	-	2.15	3.12E-02
ARHGEF37	-	-	-	-	2.15	3.14E-02
HSDL1	-	-	-	-	-2.15	3.14E-02
SIL1	-	-	-	-	-2.15	3.14E-02
MDFIC	-	-	-	-	2.15	3.17E-02
ABCC5	-	-	-	-	2.14	3.21E-02
HEY2	-	-	-	-	-2.14	3.24E-02
CLPTM1	-	-	-	-	-2.14	3.25E-02
ATOH7	-	-	-	-	-2.13	3.28E-02
GSTZ1	-	-	-	-	-2.12	3.41E-02
ATL3	-	-	-	-	-2.12	3.44E-02
EXOC4	-	-	-	-	-2.11	3.44E-02
NLRC3	-	-	-	-	-2.11	3.50E-02
LRGUK	-	-	-	-	2.11	3.51E-02
NIT2	-	-	-	-	-2.11	3.51E-02
CKS2	-	-	-	-	2.10	3.59E-02
RDH10	-	-	-	-	2.10	3.60E-02
VSTM2L	-	-	-	-	-2.10	3.61E-02
FOXD4	-	-	-	-	-2.09	3.62E-02
UFSP1	-	-	-	-	-2.09	3.62E-02
LOC100506343	-	-	-	-	-2.09	3.65E-02
PDPN	-	-	-	-	-2.09	3.67E-02
CARM1	-	-	-	-	-2.08	3.71E-02
PPOX	-	-	-	-	2.08	3.71E-02
FGD4	-	-	-	-	-2.08	3.78E-02
PIGH	-	-	-	-	2.08	3.78E-02
KIAA1033	-	-	-	-	-2.07	3.82E-02
ATP5G1	-	-	-	-	-2.06	3.89E-02
ITIH4	-	-	-	-	2.06	3.92E-02
PCM1	-	-	-	-	-2.06	3.93E-02
EMILIN2	-	-	-	-	2.06	3.95E-02
TMEM216	-	-	-	-	2.06	3.95E-02
MRAP2	-	-	-	-	-2.06	3.96E-02
CDK6	-	-	-	-	2.06	3.98E-02
TSPAN18	-	-	-	-	-2.05	3.99E-02
ALG5	-	-	-	-	-2.05	4.03E-02
ALG1L	-	-	-	-	2.05	4.05E-02
NNT	-	-	-	-	2.05	4.05E-02
AGRN	-	-	-	-	-2.05	4.06E-02
MRP63	-	-	-	-	2.05	4.08E-02
EFS	-	-	-	-	-2.04	4.13E-02
IRF6	-	-	-	-	-2.04	4.17E-02
MTCH2	-	-	-	-	-2.04	4.17E-02
ZNF429	-	-	-	-	2.04	4.18E-02
WFDC1	-	-	-	-	-2.03	4.19E-02
HMGXB3	-	-	-	-	2.03	4.24E-02

NOS2	-	-	-	-	-2.03	4.25E-02
TMEM26	-	-	-	-	-2.03	4.25E-02
ZFYVE16	-	-	-	-	-2.03	4.25E-02
ZNF197	-	-	-	-	-2.03	4.27E-02
TMEM229A	-	-	-	-	2.03	4.28E-02
SSTR1	-	-	-	-	2.02	4.31E-02
MKRN2	-	-	-	-	2.02	4.34E-02
UBE2Z	-	-	-	-	-2.02	4.37E-02
IFNW1	-	-	-	-	2.02	4.38E-02
TCFL5	-	-	-	-	2.02	4.39E-02
ITGA8	-	-	-	-	2.01	4.40E-02
COPS3	-	-	-	-	2.01	4.41E-02
ASNS	-	-	-	-	2.01	4.42E-02
G3BP1	-	-	-	-	2.00	4.51E-02
PDZRN4	-	-	-	-	2.00	4.52E-02
ALG6	-	-	-	-	-2.00	4.53E-02
COL14A1	-	-	-	-	2.00	4.58E-02
ANKRD26	-	-	-	-	-1.99	4.67E-02
KIF12	-	-	-	-	1.99	4.67E-02
DNAJA3	-	-	-	-	-1.99	4.69E-02
FAM180B	-	-	-	-	-1.99	4.70E-02
RCBTB1	-	-	-	-	-1.99	4.70E-02
TET1	-	-	-	-	-1.98	4.74E-02
DRD5	-	-	-	-	1.98	4.79E-02
PPP1R13B	-	-	-	-	1.98	4.82E-02
FAM69A	-	-	-	-	-1.97	4.83E-02
RPL23AP82	-	-	-	-	1.97	4.84E-02
SRSF12	-	-	-	-	-1.97	4.86E-02
DYNLL1	-	-	-	-	-1.97	4.88E-02
EPHA10	-	-	-	-	1.97	4.88E-02
SRR	-	-	-	-	-1.97	4.88E-02
TRIM61	-	-	-	-	-1.97	4.88E-02
C11orf49	-	-	-	-	1.97	4.89E-02
ROBO1	-	-	-	-	1.97	4.92E-02
USP48	-	-	-	-	-1.96	4.95E-02
PTK7	-	-	-	-	1.96	5.00E-02

Publications relating to this thesis

1. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. Bettencourt, C.*, Moss, D. H.*, **Flower, M.***, Wiethoff, S., Brice, A., Goizet, C., Stevanin, G., Koutsis, G., Karadima, G., Panas, M., Yescas-Gomez, P., Garcia-Velazquez, L. E., Alonso-Vilatela, M. E., Lima, M., Raposo, M., Traynor, B., Sweeney, M., Wood, N., Giunti, P., network, Spatax, Durr, A., Holmans, P.[#], Houlden, H.[#], Tabrizi, S. J.[#] and Jones, L.[#] ***Ann Neurol***, 2016 Jun;79(6):983-90. doi: 10.1002/ana.24656.
2. Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. Hensman Moss, Davina J.*, **Flower, Michael D.***, Lo, Kitty K., Miller, James R. C., van Ommen, Gert-Jan B., 't Hoen, Peter A. C., Stone, Timothy C., Guinee, Amelia, Langbehn, Douglas R., Jones, Lesley, Plagnol, Vincent, van Roon-Mom, Willeke M. C., Holmans, Peter[#] and Tabrizi, Sarah J.[#] ***Scientific Reports***, 2017 Mar 21;7:44849. doi: 10.1038/srep44849.
3. FAN1 modifies Huntington's disease progression by stabilising the expanded HTT CAG repeat. Goold, R.*, **Flower, M.***, Moss, D. H., Medway, C., Wood-Kaczmar, A., Andre, R., Farshim, P., Bates, G. P., Holmans, P., Jones, L. and Tabrizi, S. J. ***Hum Mol Genet***, 2018 Oct 24. doi: 10.1093/hmg/ddy375.
4. MSH3 modifies somatic instability and disease severity in Huntington's and myotonic dystrophy type 1. **Flower M.***, Lomeikaite V.*, Ciosi M., Cumming S., Morales F., Lo K., Hensman Moss D., Jones L., Holmans P., the TRACK-HD Investigators, the OPTIMISTIC Consortium, Monckton D.G.[#] and Tabrizi S.J.[#] ***Brain*** (accepted March 2019).

* These authors should be regarded as joint first authors.

[#] These authors jointly supervised the work.

Funding

This research was funded by:

- Medical Research Council (UK) PhD studentship (1477284).
- The Rosetrees Trust (JS16/M574).
- The European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 2012-305121 "Integrated European –omics research project for diagnosis and therapy in rare neuromuscular and neurodegenerative diseases (NEUROMICS)" [305121].
- CHDI Foundation (548613).
- UK Dementia Research Institute (DRI).

References

ACHARYA, S., WILSON, T., GRADIA, S., KANE, M. F., GUERRETTE, S., MARSISCHKY, G. T., KOLODNER, R. & FISHEL, R. 1996. hMSH2 forms specific mispair-binding complexes with hMSH3 and hMSH6. *Proc Natl Acad Sci U S A,* 93**,** 13629-34.

ACUÑA, A. I., ESPARZA, M., KRAMM, C., BELTRÁN, F. A., PARRA, A. V., CEPEDA, C., TORO, C. A., VIDAL, R. L., HETZ, C., CONCHA, I. I., BRAUCHI, S., LEVINE, M. S. & CASTRO, M. A. 2013. A failure in energy metabolism and antioxidant uptake precede symptoms of Huntington's disease in mice. *Nature Communications,* 4**,** 2917.

ADAM, R., SPIER, I., ZHAO, B., KLOTH, M., MARQUEZ, J., HINRICHSEN, I., KIRFEL, J., TAFAZZOLI, A., HORPAOPAN, S., UHLHAAS, S., STIENEN, D., FRIEDRICHS, N., ALTMULLER, J., LANER, A., HOLZAPFEL, S., PETERS, S., KAYSER, K., THIELE, H., HOLINSKI-FEDER, E., MARRA, G., KRISTIANSEN, G., NOTHEN, M. M., BUTTNER, R., MOSLEIN, G., BETZ, R. C., BRIEGER, A., LIFTON, R. P. & ARETZ, S. 2016. Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. *Am J Hum Genet,* 99**,** 337-51.

ADZHUBEI, I., JORDAN, D. M. & SUNYAEV, S. R. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet,* Chapter 7**,** Unit7 20.

AFFYMETRIX. 2016. *Affymetrix* [Online]. Available: http://www.affymetrix.com/estore/index.jsp [Accessed 21/01/2016].

AKKARI, Y. M. N., BATEMAN, R. L., REIFSTECK, C. A., OLSON, S. B. & GROMPE, M. 2000. DNA Replication Is Required To Elicit Cellular Responses to Psoralen-Induced DNA Interstrand Cross-Links. *Molecular and Cellular Biology,* 20**,** 8283-8289.

ALBERCH, J., LOPEZ, M., BADENAS, C., CARRASCO, J. L., MILA, M., MUNOZ, E. & CANALS, J. M. 2005. Association between BDNF Val66Met polymorphism and age at onset in Huntington disease. *Neurology,* 65**,** 964-5.

ALJANABI, S. M. & MARTINEZ, I. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res,* 25**,** 4692-3.

ANDERSON, S. & DEPAMPHILIS, M. L. 1979. Metabolism of Okazaki fragments during simian virus 40 DNA replication. *J Biol Chem,* 254**,** 11495-504.

ANDREW, S. E., GOLDBERG, Y. P., KREMER, B., SQUITIERI, F., THEILMANN, J., ZEISLER, J., TELENIUS, H., ADAM, S., ALMQUIST, E., ANVRET, M. & ET AL. 1994a. Huntington disease without CAG expansion: phenocopies or errors in assignment? *Am J Hum Genet,* 54**,** 852-63.

ANDREW, S. E., GOLDBERG, Y. P., KREMER, B., TELENIUS, H., THEILMANN, J., ADAM, S., STARR, E., SQUITIERI, F., LIN, B., KALCHMAN, M. A. & ET AL. 1993. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nat Genet,* 4**,** 398-403.

ANDREW, S. E., GOLDBERG, Y. P., THEILMANN, J., ZEISLER, J. & HAYDEN, M. R. 1994b. A CCG repeat polymorphism adjacent to the CAG repeat in the Huntington disease gene: implications for diagnostic accuracy and predictive testing. *Hum Mol Genet,* 3**,** 65-7.

ANTONINI, G., GRAGNANI, F., ROMANIELLO, A., PENNISI, E. M., MORINO, S., CESCHIN, V., SANTORO, L. & CRUCCU, G. 2000. Sensory involvement in spinal-bulbar muscular atrophy (Kennedy's disease). *Muscle Nerve,* 23**,** 252-8.

ANVRET, M., AHLBERG, G., GRANDELL, U., HEDBERG, B., JOHNSON, K. & EDSTROM, L. 1993. Larger expansions of the CTG repeat in muscle compared to lymphocytes from patients with myotonic dystrophy. *Hum Mol Genet,* 2**,** 1397-400.

AOKI, M., ABE, K., TOBITA, M., KAMEYA, T., WATANABE, M. & ITOYAMA, Y. 1996. Reduction of CAG expansions in cerebellar cortex and spinal cord of DRPLA. *Clin Genet,* 50**,** 199-201.

APOSTOL, B. L., SIMMONS, D. A., ZUCCATO, C., ILLES, K., PALLOS, J., CASALE, M., CONFORTI, P., RAMOS, C., ROARKE, M., KATHURIA, S., CATTANEO, E., MARSH, J. L. & THOMPSON, L. M. 2008. CEP-1347 reduces mutant huntingtin-associated neurotoxicity and restores BDNF levels in R6/2 mice. *Mol Cell Neurosci,* 39**,** 8-20.

ARBER, C., PRECIOUS, S. V., CAMBRAY, S., RISNER-JANICZEK, J. R., KELLY, C., NOAKES, Z., FJODOROVA, M., HEUER, A., UNGLESS, M. A., RODRIGUEZ, T. A., ROSSER, A. E., DUNNETT, S. B. & LI, M. 2015. Activin A directs striatal projection neuron differentiation of human pluripotent stem cells. *Development,* 142**,** 1375-86.

ARMSTRONG, J. S., KHDOUR, O. & HECHT, S. M. 2010. Does oxidative stress contribute to the pathology of Friedreich's ataxia? A radical question. *Faseb j,* 24**,** 2152-63.

ARNING, L. & EPPLEN, J. T. 2013. Genetic modifiers in Huntington's disease: fiction or fact? *Neurogenetics,* 14**,** 171-2.

ARRASATE, M., MITRA, S., SCHWEITZER, E. S., SEGAL, M. R. & FINKBEINER, S. 2004. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature,* 431**,** 805-10.

ASCHAUER, D. F., KREUZ, S. & RUMPEL, S. 2013. Analysis of transduction efficiency, tropism and axonal transport of AAV serotypes 1, 2, 5, 6, 8 and 9 in the mouse brain. *PLoS One,* 8**,** e76310.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

ASHIZAWA, T., DUBEL, J. R. & HARATI, Y. 1993. Somatic instability of CTG repeat in myotonic dystrophy. *Neurology,* 43**,** 2674-8.

ASHIZAWA, T., MONCKTON, D. G., VAISHNAV, S., PATEL, B. J., VOSKOVA, A. & CASKEY, C. T. 1996. Instability of the expanded (CTG)n repeats in the myotonin protein kinase gene in cultured lymphoblastoid cell lines from patients with myotonic dystrophy. *Genomics,* 36**,** 47-53.

ATSUTA, N., WATANABE, H., ITO, M., BANNO, H., SUZUKI, K., KATSUNO, M., TANAKA, F., TAMAKOSHI, A. & SOBUE, G. 2006. Natural history of spinal and bulbar muscular atrophy (SBMA): a study of 223 Japanese patients. *Brain,* 129**,** 1446-55.

ATWAL, R. S., DESMOND, C. R., CARON, N., MAIURI, T., XIA, J., SIPIONE, S. & TRUANT, R. 2011. Kinase inhibitors modulate huntingtin cell localization and toxicity. *Nat Chem Biol,* 7**,** 453-460.

AUBRY, L., BUGI, A., LEFORT, N., ROUSSEAU, F., PESCHANSKI, M. & PERRIER, A. L. 2008. Striatal progenitors derived from human ES cells mature into DARPP32 neurons in vitro and in quinolinic acid-lesioned rats. *Proc Natl Acad Sci U S A,* 105**,** 16707-12.

AXFORD, M. M., WANG, Y. H., NAKAMORI, M., ZANNIS-HADJOPOULOS, M., THORNTON, C. A. & PEARSON, C. E. 2013. Detection of slipped-DNAs at the trinucleotide repeats of the myotonic dystrophy type I disease locus in patient tissues. *PLoS Genet,* 9**,** e1003866.

AZIZ, N. A., JURGENS, C. K., LANDWEHRMEYER, G. B., GROUP, E. R. S., VAN ROON-MOM, W. M., VAN OMMEN, G. J., STIJNEN, T. & ROOS, R. A. 2009. Normal and mutant HTT interact to affect clinical severity and progression in Huntington disease. *Neurology,* 73**,** 1280-5.

BALLMAIER, D. & EPE, B. 2006. DNA damage by bromate: mechanism and consequences. *Toxicology,* 221**,** 166-71.

BANDELT, H. J., FORSTER, P. & ROHL, A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol,* 16**,** 37-48.

BANEZ-CORONEL, M., AYHAN, F., TARABOCHIA, A. D., ZU, T., PEREZ, B. A., TUSI, S. K., PLETNIKOVA, O., BORCHELT, D. R., ROSS, C. A., MARGOLIS, R. L., YACHNIS, A. T., TRONCOSO, J. C. & RANUM, L. P. 2015. RAN Translation in Huntington Disease. *Neuron,* 88**,** 667-77.

BATES, G., TABRIZI, S. & JONES, L. 2014. *Huntington's disease*, Oxford University Press.

BATES, G. P., DORSEY, R., GUSELLA, J. F., HAYDEN, M. R., KAY, C., LEAVITT, B. R., NANCE, M., ROSS, C. A., SCAHILL, R. I., WETZEL, R., WILD, E. J. & TABRIZI, S. J. 2015a. Huntington disease. *Nature Reviews Disease Primers*.

BATES, G. P., DORSEY, R., GUSELLA, J. F., HAYDEN, M. R., KAY, C., LEAVITT, B. R., NANCE, M., ROSS, C. A., SCAHILL, R. I., WETZEL, R., WILD, E. J. & TABRIZI, S. J. 2015b. Huntington disease. *Nat Rev Dis Primers,* 1**,** 15005.

BATES, G. P., DORSEY, R., GUSELLA, J. F., HAYDEN, M. R., KAY, C., LEAVITT, B. R., NANCE, M., ROSS, C. A., SCAHILL, R. I., WETZEL, R., WILD, E. J. & TABRIZI, S. J. 2015c. Huntington disease. *Nature Reviews Disease Primers***,** 15005.

BEAN, L. & BAYRAK-TOYDEMIR, P. 2014. American College of Medical Genetics and Genomics Standards and Guidelines for Clinical Genetics Laboratories, 2014 edition: technical standards and guidelines for Huntington disease. *Genet Med,* 16**,** e2.

BEBENEK, K. & KUNKEL, T. A. 1990. Frameshift errors initiated by nucleotide misincorporation. *Proc Natl Acad Sci U S A,* 87**,** 4946-50.

BEBENEK, K., ROBERTS, J. D. & KUNKEL, T. A. 1992. The effects of dNTP pool imbalances on frameshift fidelity during DNA replication. *J Biol Chem,* 267**,** 3589-96.

BECANOVIC, K., NORREMOLLE, A., NEAL, S. J., KAY, C., COLLINS, J. A., ARENILLAS, D., LILJA, T., GAUDENZI, G., MANOHARAN, S., DOTY, C. N., BECK, J., LAHIRI, N., PORTALES-CASAMAR, E., WARBY, S. C., CONNOLLY, C., DE SOUZA, R. A., NETWORK, R. I. O. T. E. H. S. D., TABRIZI, S. J., HERMANSON, O., LANGBEHN, D. R., HAYDEN, M. R., WASSERMAN, W. W. & LEAVITT, B. R. 2015. A SNP in the HTT promoter alters NF-kappaB binding and is a bidirectional genetic modifier of Huntington disease. *Nature Neuroscience,* 18**,** 807-16.

BECK, J., POULTER, M., HENSMAN, D., ROHRER, J. D., MAHONEY, C. J., ADAMSON, G., CAMPBELL, T., UPHILL, J., BORG, A., FRATTA, P., ORRELL, R. W., MALASPINA, A., ROWE, J., BROWN, J., HODGES, J., SIDLE, K., POLKE, J. M., HOULDEN, H., SCHOTT, J. M., FOX, N. C., ROSSOR, M. N., TABRIZI, S. J., ISAACS, A. M., HARDY, J., WARREN, J. D., COLLINGE, J. & MEAD, S. 2013. Large C9orf72 hexanucleotide repeat expansions are seen in multiple neurodegenerative syndromes and are more frequent than expected in the UK population. *Am J Hum Genet,* 92**,** 345-53.

BECONI, M. G., YATES, D., LYONS, K., MATTHEWS, K., CLIFTON, S., MEAD, T., PRIME, M., WINKLER, D., O'CONNELL, C., WALTER, D., TOLEDO-SHERMAN, L., MUNOZ-SANJUAN, I. & DOMINGUEZ, C. 2012. Metabolism and Pharmacokinetics of JM6 in Mice: JM6 Is Not a Prodrug for Ro-61-8048. *Drug Metabolism and Disposition,* 40**,** 2297-2306.

BENITEZ, J., ROBLEDO, M., RAMOS, C., AYUSO, C., ASTARLOA, R., GARCIA YEBENES, J. & BRAMBATI, B. 1995. Somatic stability in chorionic villi samples and other Huntington fetal tissues. *Hum Genet,* 96**,** 229-32.

BENN, C. L., FOX, H. & BATES, G. P. 2008a. Optimisation of region-specific reference gene selection and relative gene expression analysis methods for pre-clinical trials of Huntington's disease. *Mol Neurodegener,* 3**,** 17.

BENN, C. L., SUN, T., SADRI-VAKILI, G., MCFARLAND, K. N., DIROCCO, D. P., YOHRLING, G. J., CLARK, T. W., BOUZOU, B. & CHA, J. H. 2008b. Huntingtin modulates transcription, occupies gene promoters in vivo, and binds directly to DNA in a polyglutamine-dependent manner. *J Neurosci,* 28**,** 10720-33.

BETTENCOURT, C. & LIMA, M. 2011. Machado-Joseph Disease: from first descriptions to new perspectives. *Orphanet J Rare Dis,* 6**,** 35.

BETTENCOURT, C., MOSS, D. H., FLOWER, M., WIETHOFF, S., BRICE, A., GOIZET, C., STEVANIN, G., KOUTSIS, G., KARADIMA, G., PANAS, M., YESCAS-GOMEZ, P., GARCIA-VELAZQUEZ, L. E., ALONSO-VILATELA, M. E., LIMA, M., RAPOSO, M., TRAYNOR, B., SWEENEY, M., WOOD, N., GIUNTI, P., NETWORK, S., DURR, A., HOLMANS, P., HOULDEN, H., TABRIZI, S. J. & JONES, L. 2016. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann Neurol*.

BIDICHANDANI, S. I., PURANDARE, S. M., TAYLOR, E. E., GUMIN, G., MACHKHAS, H., HARATI, Y., GIBBS, R. A., ASHIZAWA, T. & PATEL, P. I. 1999. Somatic sequence variation at the Friedreich ataxia locus includes complete contraction of the expanded GAA triplet repeat, significant length variation in serially passaged lymphoblasts and enhanced mutagenesis in the flanking sequence. *Hum Mol Genet,* 8**,** 2425-36.

BIOCARTA 2016. www.biocarta.com.

BISCOTTI, M. A., CANAPA, A., FORCONI, M., OLMO, E. & BARUCCA, M. 2015. Transcription of tandemly repetitive DNA: functional roles. *Chromosome Res,* 23**,** 463-77.

BJORKQVIST, M., WILD, E. J., THIELE, J., SILVESTRONI, A., ANDRE, R., LAHIRI, N., RAIBON, E., LEE, R. V., BENN, C. L., SOULET, D., MAGNUSSON, A., WOODMAN, B., LANDLES, C., POULADI, M. A., HAYDEN, M. R., KHALILI-SHIRAZI, A., LOWDELL, M. W., BRUNDIN, P., BATES, G. P., LEAVITT, B. R., MOLLER, T. & TABRIZI, S. J. 2008. A novel pathogenic pathway of immune activation detectable before clinical onset in Huntington's disease. *J Exp Med,* 205**,** 1869-77.

BLEKHMAN, R., MARIONI, J. C., ZUMBO, P., STEPHENS, M. & GILAD, Y. 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res,* 20**,** 180-9.

BOGDANOV, M. B., ANDREASSEN, O. A., DEDEOGLU, A., FERRANTE, R. J. & BEAL, M. F. 2001. Increased oxidative damage to DNA in a transgenic mouse model of Huntington's disease. *J Neurochem,* 79**,** 1246-9.

BOROVECKI, F., LOVRECIC, L., ZHOU, J., JEONG, H., THEN, F., ROSAS, H. D., HERSCH, S. M., HOGARTH, P., BOUZOU, B., JENSEN, R. V. & KRAINC, D. 2005. Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc Natl Acad Sci U S A,* 102**,** 11023-8.

BORRAS, E., PINEDA, M., CADINANOS, J., DEL VALLE, J., BRIEGER, A., HINRICHSEN, I., CABANILLAS, R., NAVARRO, M., BRUNET, J., SANJUAN, X., MUSULEN, E., VAN DER KLIFT, H., LAZARO, C., PLOTZ, G., BLANCO, I. & CAPELLA, G. 2013. Refining the role of PMS2 in Lynch syndrome: germline mutational analysis improved by comprehensive assessment of variants. *J Med Genet,* 50**,** 552-63.

BOUCHARD, J., TRUONG, J., BOUCHARD, K., DUNKELBERGER, D., DESRAYAUD, S., MOUSSAOUI, S., TABRIZI, S. J., STELLA, N. & MUCHOWSKI, P. J. 2012a. Cannabinoid receptor 2 signaling in peripheral immune cells modulates disease onset and severity in mouse models of Huntington's disease. *J Neurosci,* 32**,** 18259-68.

BOUCHARD, J., TRUONG, J., BOUCHARD, K., DUNKELBERGER, D., DESRAYAUD, S., MOUSSAOUI, S., TABRIZI, S. J., STELLA, N. & MUCHOWSKI, P. J. 2012b. Cannabinoid Receptor 2 Signaling in Peripheral Immune Cells Modulates Disease Onset and Severity in Mouse Models of Huntington's Disease. *Journal of Neuroscience,* 32**,** 18259-18268.

BOUCHARD, J., TRUONG, J., BOUCHARD, K., DUNKELBERGER, D., DESRAYAUD, S., MOUSSAOUI, S., TABRIZI, S. J., STELLA, N. & MUCHOWSKI, P. J. 2012c. Cannabinoid Receptor 2 Signaling in Peripheral Immune Cells Modulates Disease Onset and Severity in Mouse Models of Huntington's Disease. *The Journal of Neuroscience,* 32**,** 18259-18268.

BOURDON, A., MINAI, L., SERRE, V., JAIS, J. P., SARZI, E., AUBERT, S., CHRETIEN, D., DE LONLAY, P., PAQUIS-FLUCKLINGER, V., ARAKAWA, H., NAKAMURA, Y., MUNNICH, A. & ROTIG, A. 2007. Mutation of RRM2B, encoding p53-controlled ribonucleotide reductase (p53R2), causes severe mitochondrial DNA depletion. *Nat Genet,* 39**,** 776-80.

BOURN, R. L., DE BIASE, I., PINTO, R. M., SANDI, C., AL-MAHDAWI, S., POOK, M. A. & BIDICHANDANI, S. I. 2012. Pms2 suppresses large expansions of the (GAA.TTC)n sequence in neuronal tissues. *PLoS One,* 7**,** e47085.

BRAINEAC. 2016. *Braineac - The Brain eQTL Almanac* [Online]. Available: http://www.braineac.org/ [Accessed 21/01/2016].

BRIERLEY, D. J. & MARTIN, S. A. 2013. Oxidative stress and the DNA mismatch repair pathway. *Antioxid Redox Signal,* 18**,** 2420-8.

BROOK, J. D., MCCURRACH, M. E., HARLEY, H. G., BUCKLER, A. J., CHURCH, D., ABURATANI, H., HUNTER, K., STANTON, V. P., THIRION, J. P., HUDSON, T. & ET AL. 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell,* 68**,** 799-808.

BROWN, M. B. 1975. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics,* 31**,** 987-992.

BROWNE, S. E., BOWLING, A. C., MACGARVEY, U., BAIK, M. J., BERGER, S. C., MUQIT, M. M., BIRD, E. D. & BEAL, M. F. 1997. Oxidative damage and metabolic dysfunction in Huntington's disease: selective vulnerability of the basal ganglia. *Ann Neurol,* 41**,** 646-53.

BUDWORTH, H., HARRIS, F. R., WILLIAMS, P., LEE DO, Y., HOLT, A., PAHNKE, J., SZCZESNY, B., ACEVEDO-TORRES, K., AYALA-PENA, S. & MCMURRAY, C. T. 2015. Suppression of Somatic Expansion Delays the Onset of Pathophysiology in a Mouse Model of Huntington's Disease. *PLoS Genet,* 11**,** e1005267.

BUDWORTH, H. & MCMURRAY, C. T. 2013. A brief history of triplet repeat diseases. *Methods Mol Biol,* 1010**,** 3-17.

BURGESS, R. C., BURMAN, B., KRUHLAK, M. J. & MISTELI, T. 2014. Activation of DNA damage response signaling by condensed chromatin. *Cell Rep,* 9**,** 1703-1717.

BURK, K., GLOBAS, C., BOSCH, S., KLOCKGETHER, T., ZUHLKE, C., DAUM, I. & DICHGANS, J. 2003. Cognitive deficits in spinocerebellar ataxia type 1, 2, and 3. *J Neurol,* 250**,** 207-11.

BURK, K., GLOBAS, C., BOSCH, S., GRABER, S., ABELE, M., BRICE, A., DICHGANS, J., DAUM, I., KLOCKGETHER T. 1999. Cognitive deficits in spinocerebellar ataxia 2. *Brain* 122**,** 769-777.

BUSH, W. S. & MOORE, J. H. 2012. Chapter 11: Genome-wide association studies. *PLoS Comput Biol,* 8**,** e1002822.

BUSSE, M. E., HUGHES, G., WILES, C. M. & ROSSER, A. E. 2008. Use of hand-held dynamometry in the evaluation of lower limb muscle strength in people with Huntington's disease. *Journal of Neurology,* 255**,** 1534-1540.

CAI, C., LANGFELDER, P., FULLER, T. F., OLDHAM, M. C., LUO, R., VAN DEN BERG, L. H., OPHOFF, R. A. & HORVATH, S. 2010. Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics,* 11**,** 589.

CALABRESE, V., LODI, R., TONON, C., D'AGATA, V., SAPIENZA, M., SCAPAGNINI, G., MANGIAMELI, A., PENNISI, G., STELLA, A. M. & BUTTERFIELD, D. A. 2005. Oxidative stress, mitochondrial dysfunction and cellular stress response in Friedreich's ataxia. *J Neurol Sci,* 233**,** 145-62.

CAMPREGHER, C., SCHMID, G., FERK, F., KNASMULLER, S., KHARE, V., KORTUM, B., DAMMANN, K., LANG, M., SCHARL, T., SPITTLER, A., ROIG, A. I., SHAY, J. W., GERNER, C. & GASCHE, C. 2012. MSH3-deficiency initiates EMAST without oncogenic transformation of human colon epithelial cells. *PLoS One,* 7**,** e50541.

CANNELLA, M., GELLERA, C., MAGLIONE, V., GIALLONARDO, P., CISLAGHI, G., MUGLIA, M., QUATTRONE, A., PIERELLI, F., DI DONATO, S. & SQUITIERI, F. 2004. The gender effect in juvenile Huntington disease patients of Italian origin. *Am J Med Genet B Neuropsychiatr Genet,* 125B**,** 92-8.

CANNELLA, M., MAGLIONE, V., MARTINO, T., RAGONA, G., FRATI, L., LI, G. M. & SQUITIERI, F. 2009. DNA instability in replicating Huntington's disease lymphoblasts. *BMC Med Genet,* 10**,** 11.

CANUGOVI, C., MISIAK, M., FERRARELLI, L. K., CROTEAU, D. L. & BOHR, V. A. 2013. The role of DNA repair in brain related disease pathology. *DNA Repair (Amst),* 12**,** 578-87.

CARROLL, J. B., BATES, G. P., STEFFAN, J., SAFT, C. & TABRIZI, S. J. 2015. Treating the whole body in Huntington's disease. *Lancet Neurol,* 14**,** 1135-42.

CARVALHO, B. S. & IRIZARRY, R. A. 2010. A framework for oligonucleotide microarray preprocessing. *Bioinformatics,* 26**,** 2363-7.

CASTEL, A. L., CLEARY, J. D. & PEARSON, C. E. 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nature Reviews Molecular Cell Biology,* 11**,** 165-170.

CATTANEO, E., ZUCCATO, C. & TARTARI, M. 2005. Normal huntingtin function: an alternative approach to Huntington's disease. *Nat Rev Neurosci,* 6**,** 919-30.

CECCALDI, R., SARANGI, P. & D'ANDREA, A. D. 2016a. The Fanconi anaemia pathway: new players and new functions. *Nat Rev Mol Cell Biol,* 17**,** 337-349.

CECCALDI, R., SARANGI, P. & D'ANDREA, A. D. 2016b. The Fanconi anaemia pathway: new players and new functions. *Nat Rev Mol Cell Biol,* 17**,** 337-49.

CHATTERJEE, N., LIN, Y., SANTILLAN, B. A., YOTNDA, P. & WILSON, J. H. 2015. Environmental stress induces trinucleotide repeat mutagenesis in human cells. *Proc Natl Acad Sci U S A,* 112**,** 3764-9.

CHATTOPADHYAY, B., GHOSH, S., GANGOPADHYAY, P. K., DAS, S. K., ROY, T., SINHA, K. K., JHA, D. K., MUKHERJEE, S. C., CHAKRABORTY, A., SINGHAL, B. S., BHATTACHARYA, A. K. & BHATTACHARYYA, N. P. 2003. Modulation of age at onset in Huntington's disease and spinocerebellar ataxia type 2 patients originated from eastern India. *Neurosci Lett,* 345**,** 93-6.

CHATURVEDI, R. K., ADHIHETTY, P., SHUKLA, S., HENNESSY, T., CALINGASAN, N., YANG, L., STARKOV, A., KIAEI, M., CANNELLA, M., SASSONE, J., CIAMMOLA, A., SQUITIERI, F. & BEAL, M. F. 2009. Impaired PGC-1alpha function in muscle in Huntington's disease. *Hum Mol Genet,* 18**,** 3048-65.

CHATURVEDI, R. K., CALINGASAN, N. Y., YANG, L., HENNESSEY, T., JOHRI, A. & BEAL, M. F. 2010. Impairment of PGC-1alpha expression, neuropathology and hepatic steatosis in a transgenic mouse model of Huntington's disease following chronic energy deprivation. *Human Molecular Genetics,* 19**,** 3190-3205.

CHAUDHURY, I., STROIK, D. R. & SOBECK, A. 2014. FANCD2-controlled chromatin access of the Fanconi-associated nuclease FAN1 is crucial for the recovery of stalled replication forks. *Mol Cell Biol,* 34**,** 3939-54.

CHEN, Y.-H., JONES, MATHEW J. K., YIN, Y., CRIST, SARAH B., COLNAGHI, L., SIMS, ROBERT J., ROTHENBERG, E., JALLEPALLI, PRASAD V. & HUANG, TONY T. 2015. ATR-Mediated Phosphorylation of FANCI Regulates Dormant Origin Firing in Response to Replication Stress. *Molecular Cell,* 58**,** 323-338.

CHRISTENSEN, J. E., NANSEN, A., MOOS, T., LU, B., GERARD, C., CHRISTENSEN, J. P. & THOMSEN, A. R. 2004. Efficient T-cell surveillance of the CNS requires expression of the CXC chemokine receptor 3. *J Neurosci,* 24**,** 4849-58.

CICCHETTI, F., LACROIX, S., CISBANI, G., VALLIERES, N., SAINT-PIERRE, M., ST-AMOUR, I., TOLOUEI, R., SKEPPER, J. N., HAUSER, R. A., MANTOVANI, D., BARKER, R. A. & FREEMAN, T. B. 2014. Mutant

huntingtin is present in neuronal grafts in Huntington disease patients. *Ann Neurol,* 76**,** 31-42.

CICCIA, A. & ELLEDGE, S. J. 2010. The DNA damage response: making it safe to play with knives. *Mol Cell,* 40**,** 179-204.

CIOSI, M., CUMMING, S. A., ALSHAMMARI, A. M., SYMEONIDI, E., HERZYK, P., MCGUINNESS, D., GALBRAITH, J., HAMILTON, G. & MONCKTON, D. G. 2018. Library preparation and MiSeq sequencing for the genotyping-by-sequencing of the Huntington disease *HTT* exon one trinucleotide repeat and the quantification of somatic mosaicism.

CLARK, A. B., DETERDING, L., TOMER, K. B. & KUNKEL, T. A. 2007a. Multiple functions for the N-terminal region of Msh6. *Nucleic Acids Res,* 35**,** 4114-23.

CLARK, A. B., VALLE, F., DROTSCHMANN, K., GARY, R. K. & KUNKEL, T. A. 2000. Functional interaction of proliferating cell nuclear antigen with MSH2-MSH6 and MSH2-MSH3 complexes. *J Biol Chem,* 275**,** 36498-501.

CLARK, R. M., DE BIASE, I., MALYKHINA, A. P., AL-MAHDAWI, S., POOK, M. & BIDICHANDANI, S. I. 2007b. The GAA triplet-repeat is unstable in the context of the human FXN locus and displays age-dependent expansions in cerebellum and DRG in a transgenic mouse model. *Hum Genet,* 120**,** 633-40.

CLEARY, J. D. & RANUM, L. P. 2014. Repeat associated non-ATG (RAN) translation: new starts in microsatellite expansion disorders. *Curr Opin Genet Dev,* 26**,** 6-15.

CLEARY, J. D., TOME, S., LOPEZ CASTEL, A., PANIGRAHI, G. B., FOIRY, L., HAGERMAN, K. A., SROKA, H., CHITAYAT, D., GOURDON, G. & PEARSON, C. E. 2010. Tissue- and age-specific DNA replication patterns at the CTG/CAG-expanded human myotonic dystrophy type 1 locus. *Nat Struct Mol Biol,* 17**,** 1079-87.

CLEMENTS, P. M., BRESLIN, C., DEEKS, E. D., BYRD, P. J., JU, L., BIEGANOWSKI, P., BRENNER, C., MOREIRA, M. C., TAYLOR, A. M. & CALDECOTT, K. W. 2004. The ataxia-oculomotor apraxia 1 gene product has a role distinct from ATM and interacts with the DNA strand break repair proteins XRCC1 and XRCC4. *DNA Repair (Amst),* 3**,** 1493-502.

CMC, C. M. C.-. 2017. *CommonMind Consortium Knowledge Portal* [Online]. Available: https://www.synapse.org//#!Synapse:syn2759792/wiki/ [Accessed 10/08/2017 2017].

COLAK, D., ZANINOVIC, N., COHEN, M. S., ROSENWAKS, Z., YANG, W. Y., GERHARDT, J., DISNEY, M. D. & JAFFREY, S. R. 2014. Promoter-bound trinucleotide repeat mRNA drives epigenetic silencing in fragile X syndrome. *Science,* 343**,** 1002-5.

COMI , G., JEFFERY , D., KAPPOS , L., MONTALBAN , X., BOYKO , A., ROCCA , M. A. & FILIPPI , M. 2012. Placebo-Controlled Trial of Oral Laquinimod for Multiple Sclerosis. *New England Journal of Medicine,* 366**,** 1000-1009.

CONFORTI, P., ZUCCATO, C., GAUDENZI, G., IERACI, A., CAMNASIO, S., BUCKLEY, N. J., MUTTI, C., COTELLI, F., CONTINI, A. & CATTANEO, E. 2013. Binding of the repressor complex REST-mSIN3b by small molecules restores neuronal gene transcription in Huntington's disease models. *J Neurochem,* 127**,** 22-35.

CONG, S. Y., PEPERS, B. A., EVERT, B. O., RUBINSZTEIN, D. C., ROOS, R. A., VAN OMMEN, G. J. & DORSMAN, J. C. 2005. Mutant huntingtin represses CBP, but not p300, by binding and protein degradation. *Mol Cell Neurosci,* 30**,** 560-71.

CONNOR, B. 2018. Concise Review: The Use of Stem Cells for Understanding and Treating Huntington's Disease. *Stem Cells,* 36**,** 146-160.

CONSORTIUM, G. O. 2016. *Gene Ontology Consortium* [Online]. Available: http://geneontology.org/ [Accessed 21/01/2016].

CONSORTIUM, H. D. I. 2017. Developmental alterations in Huntington's disease neural cells and pharmacological rescue in cells and mice. *Nat Neurosci,* 20**,** 648-660.

CONSORTIUM, H. I. 2012. Induced pluripotent stem cells from patients with Huntington's disease show CAG-repeat-expansion-associated phenotypes. *Cell Stem Cell,* 11**,** 264-78.

CONSORTIUM, S. W. G. O. T. P. G. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature,* 511**,** 421-7.

COOPER, G. M. 2000. *The Cell: A Molecular Approach*, Sunderland (MA): Sinauer Associates.

COSTA, V., ANGELINI, C., DE FEIS, I. & CICCODICOLA, A. 2010. Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol,* 2010**,** 853916.

CRAUFURD, D. & SNOWDEN, J. 2002. Neuropsychological and neuropsychiatric aspects of Huntington's disease. *Oxford Monographs on Medical Genetics,* 45**,** 62-94.

CROTTI, A. & GLASS, C. K. 2015. The choreography of neuroinflammation in Huntington's disease. *Trends Immunol,* 36**,** 364-73.

CUI, L., JEONG, H., BOROVECKI, F., PARKHURST, C. N., TANESE, N. & KRAINC, D. 2006. Transcriptional repression of PGC-1alpha by mutant huntingtin leads to mitochondrial dysfunction and neurodegeneration. *Cell,* 127**,** 59-69.

CUMMING, S. A., HAMILTON, M. J., ROBB, Y., GREGORY, H., MCWILLIAM, C., COOPER, A., ADAM, B., MCGHIE, J., HAMILTON, G., HERZYK, P., TSCHANNEN, M. R., WORTHEY, E., PETTY, R., BALLANTYNE, B., SCOTTISH MYOTONIC DYSTROPHY, C., WARNER, J., FARRUGIA, M. E., LONGMAN, C. & MONCKTON, D. G. 2018. De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. *Eur J Hum Genet*.

DALRYMPLE, A., WILD, E. J., JOUBERT, R., SATHASIVAM, K., BJORKQVIST, M., PETERSEN, A., JACKSON, G. S., ISAACS, J. D., KRISTIANSEN, M., BATES, G. P., LEAVITT, B. R., KEIR, G., WARD, M. & TABRIZI, S. J. 2007. Proteomic profiling of plasma in Huntington's disease reveals neuroinflammatory activation and biomarker candidates. *J Proteome Res,* 6**,** 2833-40.

DAVID, G., DURR, A., STEVANIN, G., CANCEL, G., ABBAS, N., BENOMAR, A., BELAL, S., LEBRE, A. S., ABADA-BENDIB, M., GRID, D., HOLMBERG, M., YAHYAOUI, M., HENTATI, F., CHKILI, T., AGID, Y. & BRICE, A. 1998. Molecular and clinical correlations in autosomal dominant cerebellar ataxia with progressive macular dystrophy (SCA7). *Hum Mol Genet,* 7**,** 165-70.

DE BIASE, I., RASMUSSEN, A., ENDRES, D., AL-MAHDAWI, S., MONTICELLI, A., COCOZZA, S., POOK, M. & BIDICHANDANI, S. I. 2007. Progressive GAA expansions in dorsal root ganglia of Friedreich's ataxia patients. *Ann Neurol,* 61**,** 55-60.

DE TEMMERMAN, N., SENECA, S., VAN STEIRTEGHEM, A., HAENTJENS, P., VAN DER ELST, J., LIEBAERS, I. & SERMON, K. D. 2008. CTG repeat instability in a human embryonic stem cell line carrying the myotonic dystrophy type 1 mutation. *Molecular Human Reproduction,* 14**,** 405-12.

DE WIND, N., DEKKER, M., CLAIJ, N., JANSEN, L., VAN KLINK, Y., RADMAN, M., RIGGINS, G., VAN DER VALK, M., VAN'T WOUT, K. & TE RIELE, H. 1999. HNPCC-like cancer predisposition in mice through simultaneous loss of Msh3 and Msh6 mismatch-repair protein functions. *Nat Genet,* 23**,** 359-62.

DEJAGER, S., BRY-GAUILLARD, H., BRUCKERT, E., EYMARD, B., SALACHAS, F., LEGUERN, E., TARDIEU, S., CHADAREVIAN, R., GIRAL, P. & TURPIN, G. 2002. A comprehensive endocrine description of Kennedy's disease revealing androgen insensitivity linked to CAG repeat length. *J Clin Endocrinol Metab,* 87**,** 3893-901.

DELLI CARRI, A., ONORATI, M., LELOS, M. J., CASTIGLIONI, V., FAEDO, A., MENON, R., CAMNASIO, S., VUONO, R., SPAIARDI, P., TALPO, F., TOSELLI, M., MARTINO, G., BARKER, R. A., DUNNETT, S. B., BIELLA, G. & CATTANEO, E. 2013. Developmentally coordinated extrinsic signals drive

human pluripotent stem cell differentiation toward authentic DARPP-32+ medium-sized spiny neurons. *Development,* 140**,** 301-12.

DELUCA, D. S., LEVIN, J. Z., SIVACHENKO, A., FENNELL, T., NAZAIRE, M. D., WILLIAMS, C., REICH, M., WINCKLER, W. & GETZ, G. 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics,* 28**,** 1530-2.

DESAI, A. & GERSON, S. 2014. Exo1 independent DNA mismatch repair involves multiple compensatory nucleases. *DNA Repair,* 21**,** 55-64.

DEVYS, D., BIANCALANA, V., ROUSSEAU, F., BOUE, J., MANDEL, J. L. & OBERLE, I. 1992. Analysis of full fragile X mutations in fetal tissues and monozygotic twins indicate that abnormal methylation and somatic heterogeneity are established early in development. *Am J Med Genet,* 43**,** 208-16.

DEYTS, C., GALAN-RODRIGUEZ, B., MARTIN, E., BOUVEYRON, N., ROZE, E., CHARVIN, D., CABOCHE, J. & BETUING, S. 2009. Dopamine D2 receptor stimulation potentiates PolyQ-Huntingtin-induced mouse striatal neuron dysfunctions via Rho/ROCK-II activation. *PLoS One,* 4**,** e8287.

DHAENENS, C. M., BURNOUF, S., SIMONIN, C., VAN BRUSSEL, E., DUHAMEL, A., DEFEBVRE, L., DURU, C., VUILLAUME, I., CAZENEUVE, C., CHARLES, P., MAISON, P., DEBRUXELLES, S., VERNY, C., GERVAIS, H., AZULAY, J. P., TRANCHANT, C., BACHOUD-LEVI, A. C., DURR, A., BUEE, L., KRYSTKOWIAK, P., SABLONNIERE, B. & BLUM, D. 2009. A genetic variation in the ADORA2A gene modifies age at onset in Huntington's disease. *Neurobiol Dis,* 35**,** 474-6.

DIFIGLIA, M., SENA-ESTEVES, M., CHASE, K., SAPP, E., PFISTER, E., SASS, M., YODER, J., REEVES, P., PANDEY, R. K., RAJEEV, K. G., MANOHARAN, M., SAH, D. W., ZAMORE, P. D. & ARONIN, N. 2007. Therapeutic silencing of mutant huntingtin with siRNA attenuates striatal and cortical neuropathology and behavioral deficits. *Proc Natl Acad Sci U S A,* 104**,** 17204-9.

DITCH, S., SAMMARCO, M. C., BANERJEE, A. & GRABCZYK, E. 2009. Progressive GAA.TTC repeat expansion in human cell lines. *PLoS Genetics,* 5**,** e1000704.

DJOUSSE, L., KNOWLTON, B., HAYDEN, M., ALMQVIST, E. W., BRINKMAN, R., ROSS, C., MARGOLIS, R., ROSENBLATT, A., DURR, A., DODE, C., MORRISON, P. J., NOVELLETTO, A., FRONTALI, M., TRENT, R. J., MCCUSKER, E., GOMEZ-TORTOSA, E., MAYO, D., JONES, R., ZANKO, A., NANCE, M., ABRAMSON, R., SUCHOWERSKY, O., PAULSEN, J., HARRISON, M., YANG, Q., CUPPLES, L. A., GUSELLA, J. F., MACDONALD, M. E. & MYERS, R. H. 2003. Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. *Am J Med Genet A,* 119A**,** 279-82.

DJOUSSE, L., KNOWLTON, B., HAYDEN, M. R., ALMQVIST, E. W., BRINKMAN, R. R., ROSS, C. A., MARGOLIS, R. L., ROSENBLATT, A., DURR, A., DODE, C., MORRISON, P. J., NOVELLETTO, A., FRONTALI, M., TRENT, R. J., MCCUSKER, E., GOMEZ-TORTOSA, E., MAYO CABRERO, D., JONES, R., ZANKO, A., NANCE, M., ABRAMSON, R. K., SUCHOWERSKY, O., PAULSEN, J. S., HARRISON, M. B., YANG, Q., CUPPLES, L. A., MYSORE, J., GUSELLA, J. F., MACDONALD, M. E. & MYERS, R. H. 2004. Evidence for a modifier of onset age in Huntington disease linked to the HD gene in 4p16. *Neurogenetics,* 5**,** 109-14.

DOLLE, M. E., GIESE, H., HOPKINS, C. L., MARTUS, H. J., HAUSDORFF, J. M. & VIJG, J. 1997. Rapid accumulation of genome rearrangements in liver but not in brain of old mice. *Nat Genet,* 17**,** 431-4.

DONATO, R., MILJAN, E. A., HINES, S. J., AOUABDI, S., POLLOCK, K., PATEL, S., EDWARDS, F. A. & SINDEN, J. D. 2007. Differential development of neuronal physiological responsiveness in two human neural stem cell lines. *BMC Neurosci,* 8**,** 36.

DORSEY, E. 2012. Characterization of a large group of individuals with huntington disease and their relatives enrolled in the COHORT study. *PLoS One,* 7**,** e29522.

DRAGATSIS, I., GOLDOWITZ, D., DEL MAR, N., DENG, Y. P., MEADE, C. A., LIU, L., SUN, Z., DIETRICH, P., YUE, J. & REINER, A. 2009. CAG repeat lengths > or =335 attenuate the phenotype in the R6/2 Huntington's disease transgenic mouse. *Neurobiol Dis,* 33**,** 315-30.

DRAGILEVA, E., HENDRICKS, A., TEED, A., GILLIS, T., LOPEZ, E. T., FRIEDBERG, E. C., KUCHERLAPATI, R., EDELMANN, W., LUNETTA, K. L., MACDONALD, M. E. & WHEELER, V. C. 2009. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiol Dis,* 33**,** 37-47.

DRIESSENS, N., VERSTEYHE, S., GHADDHAB, C., BURNIAT, A., DE DEKEN, X., VAN SANDE, J., DUMONT, J. E., MIOT, F. & CORVILAIN, B. 2009. Hydrogen peroxide induces DNA single- and double-strand breaks in thyroid cells and is therefore a potential mutagen for this organ. *Endocr Relat Cancer,* 16**,** 845-56.

DRUMMOND, J. T. 1999. Genomic amplification of the human DHFR/MSH3 locus remodels mismatch recognition and repair activities. *Adv Enzyme Regul,* 39**,** 129-41.

DRUMMOND, J. T., GENSCHEL, J., WOLF, E. & MODRICH, P. 1997. DHFR/MSH3 amplification in methotrexate-resistant cells alters the hMutSalpha/hMutSbeta ratio and reduces the efficiency of base-base mismatch repair. *Proc Natl Acad Sci U S A,* 94**,** 10144-9.

DU, J., CAMPAU, E., SORAGNI, E., JESPERSEN, C. & GOTTESFELD, J. M. 2013a. Length-dependent CTG.CAG triplet-repeat expansion in myotonic dystrophy patient-derived induced pluripotent stem cells. *Human Molecular Genetics,* 22**,** 5276-87.

DU, J., CAMPAU, E., SORAGNI, E., JESPERSEN, C. & GOTTESFELD, J. M. 2013b. Length-dependent CTG.CAG triplet-repeat expansion in myotonic dystrophy patient-derived induced pluripotent stem cells. *Hum Mol Genet,* 22**,** 5276-87.

DU, J., CAMPAU, E., SORAGNI, E., KU, S., PUCKETT, J. W., DERVAN, P. B. & GOTTESFELD, J. M. 2012a. Role of mismatch repair enzymes in GAA.TTC triplet-repeat expansion in Friedreich ataxia induced pluripotent stem cells. *Journal of Biological Chemistry,* 287**,** 29861-72.

DU, J., CAMPAU, E., SORAGNI, E., KU, S., PUCKETT, J. W., DERVAN, P. B. & GOTTESFELD, J. M. 2012b. Role of mismatch repair enzymes in GAA.TTC triplet-repeat expansion in Friedreich ataxia induced pluripotent stem cells. *J Biol Chem,* 287**,** 29861-72.

DU, Y., BALES, K. R., DODEL, R. C., LIU, X., GLINN, M. A., HORN, J. W., LITTLE, S. P. & PAUL, S. M. 1998. Alpha2-macroglobulin attenuates beta-amyloid peptide 1-40 fibril formation and associated neurotoxicity of cultured fetal rat cortical neurons. *J Neurochem,* 70**,** 1182-8.

DUDBRIDGE, F. 2013. Power and predictive accuracy of polygenic risk scores. *PLoS Genet,* 9**,** e1003348.

DUFF, K., PAULSEN, J. S., BEGLINGER, L. J., LANGBEHN, D. R., WANG, C., STOUT, J. C., ROSS, C. A., AYLWARD, E., CARLOZZI, N. E. & QUELLER, S. 2010. "Frontal" behaviors before the diagnosis of Huntington's disease and their relationship to markers of disease progression: evidence of early lack of awareness. *J Neuropsychiatry Clin Neurosci,* 22**,** 196-207.

DUGGER, B. N. & DICKSON, D. W. 2017. Pathology of Neurodegenerative Diseases. *Cold Spring Harb Perspect Biol,* 9.

DUNAH, A. W., JEONG, H., GRIFFIN, A., KIM, Y. M., STANDAERT, D. G., HERSCH, S. M., MOURADIAN, M. M., YOUNG, A. B., TANESE, N. & KRAINC, D. 2002. Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. *Science,* 296**,** 2238-43.

DURR, A. 2010. Autosomal dominant cerebellar ataxias: polyglutamine expansions and beyond. *Lancet Neurol,* 9**,** 885-94.

DURR, A., STEVANIN, G., CANCEL, G., DUYCKAERTS, C., ABBAS, N., DIDIERJEAN, O., CHNEIWEISS, H., BENOMAR, A., LYON-CAEN, O., JULIEN, J., SERDARU, M., PENET, C., AGID, Y. & BRICE, A. 1996.

Spinocerebellar ataxia 3 and Machado-Joseph disease: clinical, molecular, and neuropathological features. *Ann Neurol,* 39**,** 490-9.

DUYAO, M., AMBROSE, C., MYERS, R., NOVELLETTO, A., PERSICHETTI, F., FRONTALI, M., FOLSTEIN, S., ROSS, C., FRANZ, M., ABBOTT, M. & ET AL. 1993. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat Genet,* 4**,** 387-92.

EDELMANN, W., UMAR, A., YANG, K., HEYER, J., KUCHERLAPATI, M., LIA, M., KNEITZ, B., AVDIEVICH, E., FAN, K., WONG, E., CROUSE, G., KUNKEL, T., LIPKIN, M., KOLODNER, R. D. & KUCHERLAPATI, R. 2000. The DNA mismatch repair genes Msh3 and Msh6 cooperate in intestinal tumor suppression. *Cancer Res,* 60**,** 803-7.

EHRNHOEFER, D. E., BUTLAND, S. L., POULADI, M. A. & HAYDEN, M. R. 2009. Mouse models of Huntington disease: variations on a theme. *Dis Model Mech,* 2**,** 123-9.

EL-KHAMISY, S. F., SAIFI, G. M., WEINFELD, M., JOHANSSON, F., HELLEDAY, T., LUPSKI, J. R. & CALDECOTT, K. W. 2005. Defective DNA single-strand break repair in spinocerebellar ataxia with axonal neuropathy-1. *Nature,* 434**,** 108-13.

ELLRICHMANN, G., REICK, C., SAFT, C. & LINKER, R. A. 2013. The role of the immune system in Huntington's disease. *Clin Dev Immunol,* 2013**,** 541259.

EULALIO, A., BEHM-ANSMANT, I., SCHWEIZER, D. & IZAURRALDE, E. 2007. P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. *Mol Cell Biol,* 27**,** 3970-81.

EVANS, S. J., DOUGLAS, I., RAWLINS, M. D., WEXLER, N. S., TABRIZI, S. J. & SMEETH, L. 2013. Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records. *J Neurol Neurosurg Psychiatry,* 84**,** 1156-60.

EZZATIZADEH, V., PINTO, R. M., SANDI, C., SANDI, M., AL-MAHDAWI, S., TE RIELE, H. & POOK, M. A. 2012. The mismatch repair system protects against intergenerational GAA repeat instability in a Friedreich ataxia mouse model. *Neurobiol Dis,* 46**,** 165-71.

FAN, H. C., HO, L. I., CHI, C. S., CHEN, S. J., PENG, G. S., CHAN, T. M., LIN, S. Z. & HARN, H. J. 2014. Polyglutamine (PolyQ) diseases: genetics to treatments. *Cell Transplant,* 23**,** 441-58.

FARMER, H., MCCABE, N., LORD, C. J., TUTT, A. N., JOHNSON, D. A., RICHARDSON, T. B., SANTAROSA, M., DILLON, K. J., HICKSON, I., KNIGHTS, C., MARTIN, N. M., JACKSON, S. P., SMITH, G. C. & ASHWORTH, A. 2005. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature,* 434**,** 917-21.

FARRER, L. A. 1986. Suicide and attempted suicide in Huntington disease: implications for preclinical testing of persons at risk. *Am J Med Genet,* 24**,** 305-11.

FERRANTE, R. J., KOWALL, N. W., BEAL, M. F., RICHARDSON, E. P., JR., BIRD, E. D. & MARTIN, J. B. 1985. Selective sparing of a class of striatal neurons in Huntington's disease. *Science,* 230**,** 561-3.

FERRANTE, R. J., KUBILUS, J. K., LEE, J., RYU, H., BEESEN, A., ZUCKER, B., SMITH, K., KOWALL, N. W., RATAN, R. R., LUTHI-CARTER, R. & HERSCH, S. M. 2003. Histone deacetylase inhibition by sodium butyrate chemotherapy ameliorates the neurodegenerative phenotype in Huntington's disease mice. *J Neurosci,* 23**,** 9418-27.

FILLA, A., MARIOTTI, C., CARUSO, G., COPPOLA, G., COCOZZA, S., CASTALDO, I., CALABRESE, O., SALVATORE, E., DE MICHELE, G., RIGGIO, M. C., PAREYSON, D., GELLERA, C. & DI DONATO, S. 2000. Relative frequencies of CAG expansions in spinocerebellar ataxia and dentatorubropallidoluysian atrophy in 116 Italian families. *Eur Neurol,* 44**,** 31-6.

FINN, R. D., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., MISTRY, J., MITCHELL, A. L., POTTER, S. C., PUNTA, M., QURESHI, M., SANGRADOR-VEGAS, A., SALAZAR, G. A., TATE, J. & BATEMAN, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res,* 44**,** D279-85.

FIUMICINO, S., MARTINELLI, S., COLUSSI, C., AQUILINA, G., LEONETTI, C., CRESCENZI, M. & BIGNAMI, M. 2000. Sensitivity to DNA cross-linking chemotherapeutic agents in mismatch repair-defective cells in vitro and in xenografts. *Int J Cancer,* 85**,** 590-6.

FLORES-ROZAS, H., CLARK, D. & KOLODNER, R. D. 2000. Proliferating cell nuclear antigen and Msh2p-Msh6p interact to form an active mispair recognition complex. *Nat Genet,* 26**,** 375-8.

FOIRY, L., DONG, L., SAVOURET, C., HUBERT, L., TE RIELE, H., JUNIEN, C. & GOURDON, G. 2006. Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. *Hum Genet,* 119**,** 520-6.

FOLLONIER, C., OEHLER, J., HERRADOR, R. & LOPES, M. 2013. Friedreich's ataxia-associated GAA repeats induce replication-fork reversal and unusual molecular junctions. *Nat Struct Mol Biol,* 20**,** 486-94.

FORTUNE, M. T., VASSILOPOULOS, C., COOLBAUGH, M. I., SICILIANO, M. J. & MONCKTON, D. G. 2000. Dramatic, expansion-biased, age-dependent, tissue-specific somatic mosaicism in a transgenic mouse model of triplet repeat instability. *Hum Mol Genet,* 9**,** 439-45.

FREUDENREICH, C. H. 2018. R-loops: targets for nuclease cleavage and repeat instability. *Curr Genet,* 64**,** 789-794.

FREUDENREICH, C. H., STAVENHAGEN, J. B. & ZAKIAN, V. A. 1997. Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. *Mol Cell Biol,* 17**,** 2090-8.

FUSILLI, C., MIGLIORE, S., MAZZA, T., CONSOLI, F., DE LUCA, A., BARBAGALLO, G., CIAMMOLA, A., GATTO, E. M., CESARINI, M., ETCHEVERRY, J. L., PARISI, V., AL-ORAIMI, M., AL-HARRASI, S., AL-SALMI, Q., MARANO, M., VONSATTEL, J. G., SABATINI, U., LANDWEHRMEYER, G. B. & SQUITIERI, F. 2018. Biological and clinical manifestations of juvenile Huntington's disease: a retrospective analysis. *Lancet Neurol*.

GACY, A. M., GOELLNER, G., JURANIC, N., MACURA, S. & MCMURRAY, C. T. 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell,* 81**,** 533-40.

GANDHI, S., WOOD-KACZMAR, A., YAO, Z., PLUN-FAVREAU, H., DEAS, E., KLUPSCH, K., DOWNWARD, J., LATCHMAN, D. S., TABRIZI, S. J., WOOD, N. W., DUCHEN, M. R. & ABRAMOV, A. Y. 2009. PINK1-associated Parkinson's disease is caused by neuronal vulnerability to calcium-induced cell death. *Mol Cell,* 33**,** 627-38.

GASSET-ROSA, F., CHILLON-MARINAS, C., GOGINASHVILI, A., ATWAL, R. S., ARTATES, J. W., TABET, R., WHEELER, V. C., BANG, A. G., CLEVELAND, D. W. & LAGIER-TOURENNE, C. 2017. Polyglutamine-Expanded Huntingtin Exacerbates Age-Related Disruption of Nuclear Integrity and Nucleocytoplasmic Transport. *Neuron,* 94**,** 48-57 e4.

GAYAN, J., BROCKLEBANK, D., ANDRESEN, J. M., ALKORTA-ARANBURU, G., ZAMEEL CADER, M., ROBERTS, S. A., CHERNY, S. S., WEXLER, N. S., CARDON, L. R. & HOUSMAN, D. E. 2008. Genomewide linkage scan reveals novel loci modifying age of onset of Huntington's disease in the Venezuelan HD kindreds. *Genet Epidemiol,* 32**,** 445-53.

GE, X. Q. & BLOW, J. J. 2010. Chk1 inhibits replication factory activation but allows dormant origin firing in existing factories. *J Cell Biol,* 191**,** 1285-97.

GELLERA, C., MEONI, C., CASTELLOTTI, B., ZAPPACOSTA, B., GIROTTI, F., TARONI, F. & DIDONATO, S. 1996. Errors in Huntington disease diagnostic test caused by trinucleotide deletion in the IT15 gene. *Am J Hum Genet,* 59**,** 475-7.

GEM-HD, G. M. O. H. S. D. G.-H. C.-. 2015. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell,* 162**,** 516-26.

GERFEN, C. R. 1992. The neostriatal mosaic: multiple levels of compartmental organization in the basal ganglia. *Annu Rev Neurosci,* 15**,** 285-320.

GESCHWIND, D. H., PERLMAN, S., FIGUEROA, C. P., TREIMAN, L. J. & PULST, S. M. 1997. The prevalence and wide clinical spectrum of the spinocerebellar ataxia type 2 trinucleotide repeat in patients with autosomal dominant cerebellar ataxia. *Am J Hum Genet,* 60**,** 842-50.

GIBBS, D. L., BARATT, A., BARIC, R. S., KAWAOKA, Y., SMITH, R. D., ORWOLL, E. S., KATZE, M. G. & MCWEENEY, S. K. 2013. Protein co-expression network analysis (ProCoNA). *J Clin Bioinforma,* 3**,** 11.

GIBBS, J. R., VAN DER BRUG, M. P., HERNANDEZ, D. G., TRAYNOR, B. J., NALLS, M. A., LAI, S. L., AREPALLI, S., DILLMAN, A., RAFFERTY, I. P., TRONCOSO, J., JOHNSON, R., ZIELKE, H. R., FERRUCCI, L., LONGO, D. L., COOKSON, M. R. & SINGLETON, A. B. 2010. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet,* 6**,** e1000952.

GIUNTI, P., SABBADINI, G., SWEENEY, M. G., DAVIS, M. B., VENEZIANO, L., MANTUANO, E., FEDERICO, A., PLASMATI, R., FRONTALI, M. & WOOD, N. W. 1998. The role of the SCA2 trinucleotide repeat expansion in 89 autosomal dominant cerebellar ataxia families. Frequency, clinical and genetic correlates. *Brain,* 121 (Pt 3)**,** 459-67.

GLOBAS, C., BOSCH, S., ZUHLKE, C., DAUM, I., DICHGANS, J. & BURK, K. 2003. The cerebellum and cognition. Intellectual function in spinocerebellar ataxia type 6 (SCA6). *J Neurol,* 250**,** 1482-7.

GLOBAS, C., DU MONTCEL, S. T., BALIKO, L., BOESCH, S., DEPONDT, C., DIDONATO, S., DURR, A., FILLA, A., KLOCKGETHER, T., MARIOTTI, C., MELEGH, B., RAKOWICZ, M., RIBAI, P., ROLA, R., SCHMITZ-HUBSCH, T., SZYMANSKI, S., TIMMANN, D., VAN DE WARRENBURG, B. P., BAUER, P. & SCHOLS, L. 2008. Early symptoms in spinocerebellar ataxia type 1, 2, 3, and 6. *Mov Disord,* 23**,** 2232-8.

GOELLNER, E. M., PUTNAM, C. D. & KOLODNER, R. D. 2015. Exonuclease 1-dependent and independent mismatch repair. *DNA Repair (Amst),* 32**,** 24-32.

GOMES-PEREIRA, M. 2004. Pms2 is a genetic enhancer of trinucleotide CAG{middle dot}CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. *Human Molecular Genetics,* 13**,** 1815-1825.

GOMES-PEREIRA, M., FORTUNE, M. T., INGRAM, L., MCABNEY, J. P. & MONCKTON, D. G. 2004. Pms2 is a genetic enhancer of trinucleotide CAG.CTG repeat somatic mosaicism: implications for the mechanism of triplet repeat expansion. *Hum Mol Genet,* 13**,** 1815-25.

GOMES-PEREIRA, M., FORTUNE, M. T. & MONCKTON, D. G. 2001. Mouse tissue culture models of unstable triplet repeats: in vitro selection for larger alleles, mutational expansion bias and tissue specificity, but no association with cell division rates. *Hum Mol Genet,* 10**,** 845-54.

GOMES-PEREIRA, M., HILLEY, J. D., MORALES, F., ADAM, B., JAMES, H. E. & MONCKTON, D. G. 2014a. Disease-associated CAG.CTG triplet repeats expand rapidly in non-dividing mouse cells, but cell cycle arrest is insufficient to drive expansion. *Nucleic Acids Research,* 42**,** 7047-56.

GOMES-PEREIRA, M., HILLEY, J. D., MORALES, F., ADAM, B., JAMES, H. E. & MONCKTON, D. G. 2014b. Disease-associated CAG{middle dot}CTG triplet repeats expand rapidly in non-dividing mouse cells, but cell cycle arrest is insufficient to drive expansion. *Nucleic Acids Research,* 42**,** 7047-7056.

GOMES-PEREIRA, M. & MONCKTON, D. G. 2004a. Chemically induced increases and decreases in the rate of expansion of a CAG*CTG triplet repeat. *Nucleic Acids Res.* England.

GOMES-PEREIRA, M. & MONCKTON, D. G. 2004b. Chemically induced increases and decreases in the rate of expansion of a CAG*CTG triplet repeat. *Nucleic Acids Res,* 32**,** 2865-72.

GOMEZ-NICOLA, D., FRANSEN, N. L., SUZZI, S. & PERRY, V. H. 2013. Regulation of microglial proliferation during chronic neurodegeneration. *J Neurosci,* 33**,** 2481-93.

GONITEL, R., MOFFITT, H., SATHASIVAM, K., WOODMAN, B., DETLOFF, P. J., FAULL, R. L. & BATES, G. P. 2008. DNA instability in postmitotic neurons. *Proc Natl Acad Sci U S A,* 105**,** 3467-72.

GOOLD, R., FLOWER, M., MOSS, D. H., MEDWAY, C., WOOD-KACZMAR, A., ANDRE, R., FARSHIM, P., BATES, G. P., HOLMANS, P., JONES, L. & TABRIZI, S. J. 2018. FAN1 modifies Huntington's disease progression by stabilising the expanded HTT CAG repeat. *Hum Mol Genet*.

GOOLD, R., RABBANIAN, S., SUTTON, L., ANDRE, R., ARORA, P., MOONGA, J., CLARKE, A. R., SCHIAVO, G., JAT, P., COLLINGE, J. & TABRIZI, S. J. 2011. Rapid cell-surface prion protein conversion revealed using a novel cell system. *Nat Commun,* 2**,** 281.

GOULA, A. V., BERQUIST, B. R., WILSON, D. M., 3RD, WHEELER, V. C., TROTTIER, Y. & MERIENNE, K. 2009. Stoichiometry of base excision repair proteins correlates with increased somatic CAG instability in striatum over cerebellum in Huntington's disease transgenic mice. *PLoS Genet,* 5**,** e1000749.

GOULA, A. V., STYS, A., CHAN, J. P., TROTTIER, Y., FESTENSTEIN, R. & MERIENNE, K. 2012. Transcription elongation and tissue-specific somatic CAG instability. *PLoS Genet,* 8**,** e1003051.

GRADIA, S., ACHARYA, S. & FISHEL, R. 1997. The human mismatch recognition complex hMSH2-hMSH6 functions as a novel molecular switch. *Cell,* 91**,** 995-1005.

GREGORY, S., SCAHILL, R. I., REES, G. & TABRIZI, S. 2018. Magnetic Resonance Imaging in Huntington's Disease. *Methods Mol Biol,* 1780**,** 303-328.

GRIMA, J. C., DAIGLE, J. G., ARBEZ, N., CUNNINGHAM, K. C., ZHANG, K., OCHABA, J., GEATER, C., MOROZKO, E., STOCKSDALE, J., GLATZER, J. C., PHAM, J. T., AHMED, I., PENG, Q., WADHWA, H., PLETNIKOVA, O., TRONCOSO, J. C., DUAN, W., SNYDER, S. H., RANUM, L. P. W., THOMPSON, L. M., LLOYD, T. E., ROSS, C. A. & ROTHSTEIN, J. D. 2017. Mutant Huntingtin Disrupts the Nuclear Pore Complex. *Neuron,* 94**,** 93-107 e6.

GROUP, H. S. 1996. Unified Huntington's Disease Rating Scale: reliability and consistency. Huntington Study Group. *Mov Disord,* 11**,** 136-42.

GROUP, T. H. S. D. C. R. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell,* 72**,** 971-83.

GTEX, G. T. C.-. 2015. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science,* 348**,** 648-60.

GULBIS, J. M., KELMAN, Z., HURWITZ, J., O'DONNELL, M. & KURIYAN, J. 1996. Structure of the C-terminal region of p21(WAF1/CIP1) complexed with human PCNA. *Cell,* 87**,** 297-306.

GUPTA, S., GELLERT, M. & YANG, W. 2011a. Mechanism of mismatch recognition revealed by human MutSbeta bound to unpaired DNA loops. *Nat Struct Mol Biol,* 19**,** 72-8.

GUPTA, S., GELLERT, M. & YANG, W. 2011b. Mechanism of mismatch recognition revealed by human MutSβ bound to unpaired DNA loops. *Nat Struct Mol Biol,* 19**,** 72-78.

GUSELLA, J. F., MACDONALD, M. E. & LEE, J. M. 2014. Genetic modifiers of Huntington's disease. *Mov Disord,* 29**,** 1359-65.

GUSELLA, J. F., WEXLER, N. S., CONNEALLY, P. M., NAYLOR, S. L., ANDERSON, M. A., TANZI, R. E., WATKINS, P. C., OTTINA, K., WALLACE, M. R., SAKAGUCHI, A. Y. & ET AL. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature,* 306**,** 234-8.

GUSEV, A., KO, A., SHI, H., BHATIA, G., CHUNG, W., PENNINX, B. W., JANSEN, R., DE GEUS, E. J., BOOMSMA, D. I., WRIGHT, F. A., SULLIVAN, P. F., NIKKOLA, E., ALVAREZ, M., CIVELEK, M., LUSIS, A. J., LEHTIMAKI, T., RAITOHARJU, E., KAHONEN, M., SEPPALA, I., RAITAKARI, O. T., KUUSISTO, J., LAAKSO, M., PRICE, A. L., PAJUKANTA, P. & PASANIUC, B. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet,* 48**,** 245-52.

GWON, G. H., KIM, Y., LIU, Y., WATSON, A. T., JO, A., ETHERIDGE, T. J., YUAN, F., ZHANG, Y., CARR, A. M. & CHO, Y. 2014. Crystal structure of a Fanconi anemia-associated nuclease homolog bound to 5' flap DNA: basis of interstrand cross-link repair by FAN1. *Genes Dev,* 28**,** 2276-90.

HALABI, A., DITCH, S., WANG, J. & GRABCZYK, E. 2012a. DNA mismatch repair complex MutSbeta promotes GAA.TTC repeat expansion in human cells. *Journal of Biological Chemistry,* 287**,** 29958-67.

HALABI, A., DITCH, S., WANG, J. & GRABCZYK, E. 2012b. DNA mismatch repair complex MutSbeta promotes GAA.TTC repeat expansion in human cells. *J Biol Chem,* 287**,** 29958-67.

HALL, A. C., OSTROWSKI, L. A., PIETROBON, V. & MEKHAIL, K. 2017. Repetitive DNA loci and their modulation by the non-canonical nucleic acid structures R-loops and G-quadruplexes. *Nucleus,* 8**,** 162-181.

HANADA, K., BUDZOWSKA, M., DAVIES, S. L., VAN DRUNEN, E., ONIZAWA, H., BEVERLOO, H. B., MAAS, A., ESSERS, J., HICKSON, I. D. & KANAAR, R. 2007. The structure-specific endonuclease Mus81 contributes to replication restart by generating double-strand DNA breaks. *Nat Struct Mol Biol,* 14**,** 1096-104.

HANADA, K., BUDZOWSKA, M., MODESTI, M., MAAS, A., WYMAN, C., ESSERS, J. & KANAAR, R. 2006. The structure-specific endonuclease Mus81-Eme1 promotes conversion of interstrand DNA crosslinks into double-strands breaks. *EMBO J,* 25**,** 4921-32.

HARPER, P. S. 2001. *Myotonic Dystrophy. 3rd edn London*, Saunders WB.

HASHIDA, H., GOTO, J., KURISAKI, H., MIZUSAWA, H. & KANAZAWA, I. 1997. Brain regional differences in the expansion of a CAG repeat in the spinocerebellar ataxias: dentatorubral-pallidoluysian atrophy, Machado-Joseph disease, and spinocerebellar ataxia type 1. *Ann Neurol,* 41**,** 505-11.

HASHIDA, H., GOTO, J., SUZUKI, T., JEONG, S., MASUDA, N., OOIE, T., TACHIIRI, Y., TSUCHIYA, H. & KANAZAWA, I. 2001. Single cell analysis of CAG repeat in brains of dentatorubral-pallidoluysian atrophy (DRPLA). *J Neurol Sci,* 190**,** 87-93.

HAUGEN, A. C., GOEL, A., YAMADA, K., MARRA, G., NGUYEN, T. P., NAGASAKA, T., KANAZAWA, S., KOIKE, J., KIKUCHI, Y., ZHONG, X., ARITA, M., SHIBUYA, K., OSHIMURA, M., HEMMI, H., BOLAND, C. R. & KOI, M. 2008. Genetic instability caused by loss of MutS homologue 3 in human colorectal cancer. *Cancer Res,* 68**,** 8465-72.

HAY, R. T. & DEPAMPHILIS, M. L. 1982. Initiation of SV40 DNA replication in vivo: location and structure of 5' ends of DNA synthesized in the ori region. *Cell,* 28**,** 767-79.

HAYES, S., TURECKI, G., BRISEBOIS, K., LOPES-CENDES, I., GASPAR, C., RIESS, O., RANUM, L. P., PULST, S. M. & ROULEAU, G. A. 2000. CAG repeat length in RAI1 is associated with age at onset variability in spinocerebellar ataxia type 2 (SCA2). *Hum Mol Genet,* 9**,** 1753-8.

HAZEKI, N., TSUKAMOTO, T., YAZAWA, I., KOYAMA, M., HATTORI, S., SOMEKI, I., IWATSUBO, T., NAKAMURA, K., GOTO, J. & KANAZAWA, I. 2002. Ultrastructure of nuclear aggregates formed by expressing an expanded polyglutamine. *Biochem Biophys Res Commun,* 294**,** 429-40.

HENEKA, M. T., KUMMER, M. P. & LATZ, E. 2014. Innate immune activation in neurodegenerative disease. *Nat Rev Immunol,* 14**,** 463-77.

HENSMAN MOSS, D. J., FLOWER, M. D., LO, K. K., MILLER, J. R. C., VAN OMMEN, G.-J. B., 'T HOEN, P. A. C., STONE, T. C., GUINEE, A., LANGBEHN, D. R., JONES, L., PLAGNOL, V., VAN ROON-MOM, W. M. C., HOLMANS, P. & TABRIZI, S. J. 2017a. Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. *Scientific Reports,* 7**,** 44849.

HENSMAN MOSS, D. J. H., PARDINAS, A. F., LANGBEHN, D., LO, K., LEAVITT, B. R., ROOS, R., DURR, A., MEAD, S., INVESTIGATORS, T.-H., INVESTIGATORS, R., HOLMANS, P., JONES, L. & TABRIZI, S. J.

2017b. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurol*.

HOCKLY, E., RICHON, V. M., WOODMAN, B., SMITH, D. L., ZHOU, X., ROSA, E., SATHASIVAM, K., GHAZI-NOORI, S., MAHAL, A., LOWDEN, P. A., STEFFAN, J. S., MARSH, J. L., THOMPSON, L. M., LEWIS, C. M., MARKS, P. A. & BATES, G. P. 2003. Suberoylanilide hydroxamic acid, a histone deacetylase inhibitor, ameliorates motor deficits in a mouse model of Huntington's disease. *Proc Natl Acad Sci U S A,* 100**,** 2041-6.

HODGES, A. 2006. Regional and cellular gene expression changes in human Huntington's disease brain. *Human Molecular Genetics,* 15**,** 965-977.

HODGES, A., STRAND, A. D., ARAGAKI, A. K., KUHN, A., SENGSTAG, T., HUGHES, G., ELLISTON, L. A., HARTOG, C., GOLDSTEIN, D. R., THU, D., HOLLINGSWORTH, Z. R., COLLIN, F., SYNEK, B., HOLMANS, P. A., YOUNG, A. B., WEXLER, N. S., DELORENZI, M., KOOPERBERG, C., AUGOOD, S. J., FAULL, R. L., OLSON, J. M., JONES, L. & LUTHI-CARTER, R. 2006. Regional and cellular gene expression changes in human Huntington's disease brain. *Hum Mol Genet,* 15**,** 965-77.

HOFFROGGE, R., BEYER, S., VOLKER, U., UHRMACHER, A. M. & ROLFS, A. 2006. 2-DE proteomic profiling of neuronal stem cells. *Neurodegener Dis,* 3**,** 112-21.

HOLMANS, P. A., MASSEY, T. H. & JONES, L. 2017. Genetic modifiers of Mendelian disease: Huntington's disease and the trinucleotide repeat disorders. *Hum Mol Genet,* 26**,** R83-r90.

HONG, S., BEJA-GLASSER, V. F., NFONOYIM, B. M., FROUIN, A., LI, S., RAMAKRISHNAN, S., MERRY, K. M., SHI, Q., ROSENTHAL, A., BARRES, B. A., LEMERE, C. A., SELKOE, D. J. & STEVENS, B. 2016a. Complement and microglia mediate early synapse loss in Alzheimer mouse models. *Science*.

HONG, S., DISSING-OLESEN, L. & STEVENS, B. 2016b. New insights on the role of microglia in synaptic pruning in health and disease. *Curr Opin Neurobiol,* 36**,** 128-34.

HORVATH, S., ZHANG, Y., LANGFELDER, P., KAHN, R. S., BOKS, M. P., VAN EIJK, K., VAN DEN BERG, L. H. & OPHOFF, R. A. 2012. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol,* 13**,** R97.

HUANG, J., LIU, S., BELLANI, M. A., THAZHATHVEETIL, A. K., LING, C., DE WINTER, J. P., WANG, Y., WANG, W. & SEIDMAN, M. M. 2013. The DNA translocase FANCM/MHF promotes replication traverse of DNA interstrand crosslinks. *Mol Cell,* 52**,** 434-46.

HUANG, M. & D'ANDREA, A. D. 2010. A new nuclease member of the FAN club. *Nat Struct Mol Biol,* 17**,** 926-8.

HUGHES, A. J., L. 2014. Pathogenic Mechanisms. *In:* BATES, G. P., TABRIZI, S.J., JONES, L (ed.) *Huntington's disease.* 4th ed. Oxford: OUP.

HUNTER, A., TSILFIDIS, C., METTLER, G., JACOB, P., MAHADEVAN, M., SURH, L. & KORNELUK, R. 1992. The correlation of age of onset with CTG trinucleotide repeat amplification in myotonic dystrophy. *J Med Genet,* 29**,** 774-9.

HUNTINGTON, G. 1872. On chorea. The Medical and Surgical Reporter of Philadelphia. *cited in H. Skirton (2005),'Huntington's Disease: a nursing perspective', MedSurg Nursing,* 14**,** 167-174.

IKEUCHI, T., KOIDE, R., TANAKA, H., ONODERA, O., IGARASHI, S., TAKAHASHI, H., KONDO, R., ISHIKAWA, A., TOMODA, A., MIIKE, T. & ET AL. 1995. Dentatorubral-pallidoluysian atrophy: clinical features are closely related to unstable expansions of trinucleotide (CAG) repeat. *Ann Neurol,* 37**,** 769-75.

ILLUMINA. 2014. *TruSeq(R) RNA Sample Preparation v2 Guide* [Online]. Available: http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqrna/truseq-rna-sample-prep-v2-guide-15026495-f.pdf [Accessed 12/01/2016].

INSTITUTE, N. C. 2012. *Home : Pathway Interaction Database* [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/ [Accessed 21/01/2016].

INTERNATIONAL GENOMICS OF ALZHEIMER'S DISEASE, C. 2015. Convergent genetic and expression data implicate immunity in Alzheimer's disease. *Alzheimers Dement,* 11**,** 658-71.

IONITA-LAZA, I., XU, B., MAKAROV, V., BUXBAUM, J. D., ROOS, J. L., GOGOS, J. A. & KARAYIORGOU, M. 2014. Scan statistic-based analysis of exome sequencing data identifies FAN1 at 15q13.3 as a susceptibility gene for schizophrenia and autism. *Proc Natl Acad Sci U S A,* 111**,** 343-8.

IRVING, J. A. & HALL, A. G. 2001. Mismatch repair defects as a cause of resistance to cytotoxic drugs. *Expert Rev Anticancer Ther,* 1**,** 149-58.

ISHIGURO, H., YAMADA, K., SAWADA, H., NISHII, K., ICHINO, N., SAWADA, M., KUROSAWA, Y., MATSUSHITA, N., KOBAYASHI, K., GOTO, J., HASHIDA, H., MASUDA, N., KANAZAWA, I. & NAGATSU, T. 2001. Age-dependent and tissue-specific CAG repeat instability occurs in mouse knock-in for a mutant Huntington's disease gene. *J Neurosci Res,* 65**,** 289-97.

IYER, R. R., PLUCIENNIK, A., NAPIERALA, M. & WELLS, R. D. 2015. DNA triplet repeat expansion and mismatch repair. *Annu Rev Biochem,* 84**,** 199-226.

JACKSON, S. E. & CHESTER, J. D. 2015. Personalised cancer medicine. *Int J Cancer,* 137**,** 262-6.

JACKSON, S. P. 2002. Sensing and repairing DNA double-strand breaks. *Carcinogenesis,* 23**,** 687-96.

JACQUET, L., NEUEDER, A., FOLDES, G., KARAGIANNIS, P., HOBBS, C., JOLINON, N., MIOULANE, M., SAKAI, T., HARDING, S. E. & ILIC, D. 2015. Three Huntington's Disease Specific Mutation-Carrying Human Embryonic Stem Cell Lines Have Stable Number of CAG Repeats upon In Vitro Differentiation into Cardiomyocytes. *PLoS One,* 10**,** e0126860.

JAMA, M., MILLSON, A., MILLER, C. E. & LYON, E. 2013. Triplet repeat primed PCR simplifies testing for Huntington disease. *J Mol Diagn,* 15**,** 255-62.

JEDELE, K. B., WAHL, D., CHAHROKH-ZADEH, S., WIRTZ, A., MURKEN, J. & HOLINSKI-FEDER, E. 1998. Spinal and bulbar muscular atrophy (SBMA): somatic stability of an expanded CAG repeat in fetal tissues. *Clin Genet,* 54**,** 148-51.

JIANG, M., PENG, Q., LIU, X., JIN, J., HOU, Z., ZHANG, J., MORI, S., ROSS, C. A., YE, K. & DUAN, W. 2013. Small-molecule TrkB receptor agonists improve motor function and extend survival in a mouse model of Huntington's disease. *Hum Mol Genet,* 22**,** 2462-70.

JIN, H. & CHO, Y. 2017. Structural and functional relationships of FAN1. *DNA Repair (Amst)*.

JIN, J., ALBERTZ, J., GUO, Z., PENG, Q., RUDOW, G., TRONCOSO, J. C., ROSS, C. A. & DUAN, W. 2013. Neuroprotective effects of PPAR-gamma agonist rosiglitazone in N171-82Q mouse model of Huntington's disease. *J Neurochem,* 125**,** 410-9.

JIRICNY, J. 2000. Mismatch repair: the praying hands of fidelity. *Curr Biol,* 10**,** R788-90.

JIRICNY, J. 2006. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol,* 7**,** 335-46.

JODEIRI FARSHBAF, M. & GHAEDI, K. 2017. Huntington's Disease and Mitochondria. *Neurotox Res,* 32**,** 518-529.

JOHNSON, R. & BUCKLEY, N. J. 2009. Gene dysregulation in Huntington's disease: REST, microRNAs and beyond. *Neuromolecular Med,* 11**,** 183-99.

JOHNSON, R., ZUCCATO, C., BELYAEV, N. D., GUEST, D. J., CATTANEO, E. & BUCKLEY, N. J. 2008. A microRNA-based gene dysregulation pathway in Huntington's disease. *Neurobiol Dis,* 29**,** 438-45.

JOHRI, A., CHANDRA, A. & BEAL, M. F. 2013. PGC-1alpha, mitochondrial dysfunction, and Huntington's disease. *Free Radic Biol Med,* 62**,** 37-46.

JONES, L., HOULDEN, H. & TABRIZI, S. J. 2017. DNA repair in the trinucleotide repeat disorders. *Lancet Neurol,* 16**,** 88-96.

JONSON, I., OUGLAND, R., KLUNGLAND, A. & LARSEN, E. 2013a. Oxidative stress causes DNA triplet expansion in Huntington's disease mouse embryonic stem cells. *Stem Cell Res,* 11**,** 1264-71.

JONSON, I., OUGLAND, R. & LARSEN, E. 2013b. DNA repair mechanisms in Huntington's disease. *Mol Neurobiol,* 47**,** 1093-102.

KADYROV, F. A., DZANTIEV, L., CONSTANTIN, N. & MODRICH, P. 2006. Endonucleolytic function of MutLalpha in human mismatch repair. *Cell,* 126**,** 297-308.

KALLBERG, M., WANG, H., WANG, S., PENG, J., WANG, Z., LU, H. & XU, J. 2012. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc,* 7**,** 1511-22.

KANG, S., JAWORSKI, A., OHSHIMA, K. & WELLS, R. D. 1995. Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in E. coli. *Nat Genet,* 10**,** 213-8.

KANTARTZIS, A., WILLIAMS, G. M., BALAKRISHNAN, L., ROBERTS, R. L., SURTEES, J. A. & BAMBARA, R. A. 2012. Msh2-Msh3 interferes with Okazaki fragment processing to promote trinucleotide repeat expansions. *Cell Rep,* 2**,** 216-22.

KAPLAN, S., ITZKOVITZ, S. & SHAPIRO, E. 2007. A universal mechanism ties genotype to phenotype in trinucleotide diseases. *PLoS Comput Biol,* 3.

KARRAN, P. & ATTARD, N. 2008. Thiopurines in current medical practice: molecular mechanisms and contributions to therapy-related cancer. *Nat Rev Cancer,* 8**,** 24-36.

KAWAI, Y., SUENAGA, M., WATANABE, H., ITO, M., KATO, K., KATO, T., ITO, K., TANAKA, F. & SOBUE, G. 2008. Prefrontal hypoperfusion and cognitive dysfunction correlates in spinocerebellar ataxia type 6. *J Neurol Sci,* 271**,** 68-74.

KAWAI, Y., SUENAGA, M., WATANABE, H. & SOBUE, G. 2009. Cognitive impairment in spinocerebellar degeneration. *Eur Neurol,* 61**,** 257-68.

KEE, Y. & D'ANDREA, A. D. 2010. Expanded roles of the Fanconi anemia pathway in preserving genomic stability. *Genes Dev,* 24**,** 1680-94.

KEGEL-GLEASON, K. B. 2013. Huntingtin interactions with membrane phospholipids: strategic targets for therapeutic intervention? *J Huntingtons Dis,* 2**,** 239-50.

KEGG. 2016. *KEGG: Kyoto Encyclopedia of Genes and Genomes* [Online]. Available: http://www.kegg.jp/ [Accessed 21/01/2016].

KENNEDY, L., EVANS, E., CHEN, C. M., CRAVEN, L., DETLOFF, P. J., ENNIS, M. & SHELBOURNE, P. F. 2003. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum Mol Genet,* 12**,** 3359-67.

KENNEDY, L. & SHELBOURNE, P. F. 2000. Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? *Hum Mol Genet,* 9**,** 2539-44.

KENNEDY, W. R., ALTER, M. & SUNG, J. H. 1968. Progressive proximal spinal and bulbar muscular atrophy of late onset. A sex-linked recessive trait. *Neurology,* 18**,** 671-80.

KEOGH, N., CHAN, K. Y., LI, G. M. & LAHUE, R. S. 2017. MutSbeta abundance and Msh3 ATP hydrolysis activity are important drivers of CTG CAG repeat expansions. *Nucleic Acids Research,* 45**,** 10068-10078.

KEUM, J. W., SHIN, A., GILLIS, T., MYSORE, J. S., ABU ELNEEL, K., LUCENTE, D., HADZI, T., HOLMANS, P., JONES, L., ORTH, M., KWAK, S., MACDONALD, M. E., GUSELLA, J. F. & LEE, J. M. 2016. The HTT CAG-Expansion Mutation Determines Age at Death but Not Disease Duration in Huntington Disease. *Am J Hum Genet,* 98**,** 287-98.

KHAN, N., KOLIMI, N. & RATHINAVELAN, T. 2015. Twisting right to left: A...A mismatch in a CAG trinucleotide repeat overexpansion provokes left-handed Z-DNA conformation. *PLoS Comput Biol,* 11**,** e1004162.

KHOSHNAN, A., KO, J., WATKIN, E. E., PAIGE, L. A., REINHART, P. H. & PATTERSON, P. H. 2004. Activation of the IkappaB kinase complex and nuclear factor-kappaB contributes to mutant huntingtin neurotoxicity. *J Neurosci,* 24**,** 7999-8008.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol,* 14**,** R36.

KIM, H. & D'ANDREA, A. D. 2012. Regulation of DNA cross-link repair by the Fanconi anemia/BRCA pathway. *Genes Dev,* 26**,** 1393-408.

KIM, M., LEE, H. S., LAFORET, G., MCINTYRE, C., MARTIN, E. J., CHANG, P., KIM, T. W., WILLIAMS, M., REDDY, P. H., TAGLE, D., BOYCE, F. M., WON, L., HELLER, A., ARONIN, N. & DIFIGLIA, M. 1999. Mutant huntingtin expression in clonal striatal cells: dissociation of inclusion formation and neuronal survival by caspase inhibition. *J Neurosci,* 19**,** 964-73.

KLECZKOWSKA, H. E., MARRA, G., LETTIERI, T. & JIRICNY, J. 2001. hMSH3 and hMSH6 interact with PCNA and colocalize with it to replication foci. *Genes Dev,* 15**,** 724-36.

KNOWLES, T. P., VENDRUSCOLO, M. & DOBSON, C. M. 2014. The amyloid state and its association with protein misfolding diseases. *Nat Rev Mol Cell Biol,* 15**,** 384-96.

KOVALENKO, M., DRAGILEVA, E., ST CLAIRE, J., GILLIS, T., GUIDE, J. R., NEW, J., DONG, H., KUCHERLAPATI, R., KUCHERLAPATI, M. H., EHRLICH, M. E., LEE, J. M. & WHEELER, V. C. 2012. Msh2 acts in medium-spiny striatal neurons as an enhancer of CAG instability and mutant huntingtin phenotypes in Huntington's disease knock-in mice. *PLoS One,* 7**,** e44273.

KOVTUN, I. V., LIU, Y., BJORAS, M., KLUNGLAND, A., WILSON, S. H. & MCMURRAY, C. T. 2007. OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. *Nature,* 447**,** 447-52.

KOVTUN, I. V. & MCMURRAY, C. T. 2001. Trinucleotide expansion in haploid germ cells by gap repair. *Nat Genet,* 27**,** 407-11.

KOVTUN, I. V., THORNHILL, A. R. & MCMURRAY, C. T. 2004. Somatic deletion events occur during early embryonic development and modify the extent of CAG expansion in subsequent generations. *Hum Mol Genet,* 13**,** 3057-68.

KRASILNIKOVA, M. M. & MIRKIN, S. M. 2004. Replication stalling at Friedreich's ataxia (GAA)n repeats in vivo. *Mol Cell Biol,* 24**,** 2286-95.

KRATZ, K., SCHOPF, B., KADEN, S., SENDOEL, A., EBERHARD, R., LADEMANN, C., CANNAVO, E., SARTORI, A. A., HENGARTNER, M. O. & JIRICNY, J. 2010a. Deficiency of FANCD2-associated nuclease KIAA1018/FAN1 sensitizes cells to interstrand crosslinking agents. *Cell,* 142**,** 77-88.

KRATZ, K., SCHOPF, B., KADEN, S., SENDOEL, A., EBERHARD, R., LADEMANN, C., CANNAVO, E., SARTORI, A. A., HENGARTNER, M. O. & JIRICNY, J. 2010b. Deficiency of FANCD2-associated nuclease KIAA1018/FAN1 sensitizes cells to interstrand crosslinking agents. *Cell.* United States: 2010 Elsevier Inc.

KRAUS-PERROTTA, C. & LAGALWAR, S. 2016. Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. *Cerebellum Ataxias,* 3**,** 20.

KU, S., SORAGNI, E., CAMPAU, E., THOMAS, E. A., ALTUN, G., LAURENT, L. C., LORING, J. F., NAPIERALA, M. & GOTTESFELD, J. M. 2010. Friedreich's ataxia induced pluripotent stem cells model intergenerational GAATTC triplet repeat instability. *Cell Stem Cell,* 7**,** 631-7.

KUMAR, A., NARAYANAN, K., CHAUDHARY, R. K., MISHRA, S., KUMAR, S., VINOTH, K. J., PADMANABHAN, P. & GULYAS, B. 2016. Current Perspective of Stem Cell Therapy in Neurodegenerative and Metabolic Diseases. *Mol Neurobiol*.

KUMAR, P., HENIKOFF, S. & NG, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc,* 4**,** 1073-81.

KUNKEL, T. A. & ERIE, D. A. 2005. DNA mismatch repair. *Annu Rev Biochem,* 74**,** 681-710.

KUO, M. L., SY, A. J., XUE, L., CHI, M., LEE, M. T., YEN, T., CHIANG, M. I., CHANG, L., CHU, P. & YEN, Y. 2012. RRM2B suppresses activation of the oxidative stress pathway and is up-regulated by p53 during senescence. *Sci Rep,* 2**,** 822.

KURASHIGE, T., SHIMAMURA, M. & NAGAYAMA, Y. 2016. Differences in quantification of DNA double-strand breaks assessed by 53BP1/gammaH2AX focus formation assays and the comet assay in mammalian cells treated with irradiation and N-acetyl-L-cysteine. *J Radiat Res,* 57**,** 312-7.

KURIHARA, T., WARR, G., LOY, J. & BRAVO, R. 1997. Defects in macrophage recruitment and host defense in mice lacking the CCR2 chemokine receptor. *J Exp Med,* 186**,** 1757-62.

KWAN, W., MAGNUSSON, A., CHOU, A., ADAME, A., CARSON, M. J., KOHSAKA, S., MASLIAH, E., MOLLER, T., RANSOHOFF, R., TABRIZI, S. J., BJORKQVIST, M. & MUCHOWSKI, P. J. 2012a. Bone Marrow Transplantation Confers Modest Benefits in Mouse Models of Huntington's Disease. *Journal of Neuroscience,* 32**,** 133-142.

KWAN, W., TRAGER, U., DAVALOS, D., CHOU, A., BOUCHARD, J., ANDRE, R., MILLER, A., WEISS, A., GIORGINI, F., CHEAH, C., MOLLER, T., STELLA, N., AKASSOGLOU, K., TABRIZI, S. J. & MUCHOWSKI, P. J. 2012b. Mutant huntingtin impairs immune cell migration in Huntington disease. *J Clin Invest,* 122**,** 4737-47.

KWAN, W., TRÄGER, U., DAVALOS, D., CHOU, A., BOUCHARD, J., ANDRE, R., MILLER, A., WEISS, A., GIORGINI, F., CHEAH, C., MÖLLER, T., STELLA, N., AKASSOGLOU, K., TABRIZI, S. J. & MUCHOWSKI, P. J. 2012c. Mutant huntingtin impairs immune cell migration in Huntington disease. *Journal of Clinical Investigation,* 122**,** 4737-4747.

LA SPADA, A. R. 1997. Trinucleotide repeat instability: genetic features and molecular mechanisms. *Brain Pathol,* 7**,** 943-63.

LABADORF, A., HOSS, A. G., LAGOMARSINO, V., LATOURELLE, J. C., HADZI, T. C., BREGU, J., MACDONALD, M. E., GUSELLA, J. F., CHEN, J.-F., AKBARIAN, S., WENG, Z. & MYERS, R. H. 2015. RNA Sequence Analysis of Human Huntington Disease Brain Reveals an Extensive Increase in Inflammatory and Developmental Gene Expression. *PLOS ONE,* 10**,** e0143563.

LABBADIA, J. & MORIMOTO, R. I. 2013. Huntington's disease: underlying molecular mechanisms and emerging concepts. *Trends Biochem Sci,* 38**,** 378-85.

LABBADIA, J., NOVOSELOV, S. S., BETT, J. S., WEISS, A., PAGANETTI, P., BATES, G. P. & CHEETHAM, M. E. 2012. Suppression of protein aggregation by chaperone modification of high molecular weight complexes. *Brain,* 135**,** 1180-1196.

LACAZETTE, E. 2017. A laboratory practical illustrating the use of the ChIP-qPCR method in a robust model: Estrogen receptor alpha immunoprecipitation using Mcf-7 culture cells. *Biochem Mol Biol Educ,* 45**,** 152-160.

LACHAUD, C., MORENO, A., MARCHESI, F., TOTH, R., BLOW, J. J. & ROUSE, J. 2016a. Ubiquitinated Fancd2 recruits Fan1 to stalled replication forks to prevent genome instability. *Science*.

LACHAUD, C., SLEAN, M., MARCHESI, F., LOCK, C., ODELL, E., CASTOR, D., TOTH, R. & ROUSE, J. 2016b. Karyomegalic interstitial nephritis and DNA damage-induced polyploidy in Fan1 nuclease-defective knock-in mice. *Genes Dev,* 30**,** 639-44.

LAI, Y., BEAVER, J. M., LORENTE, K., MELO, J., RAMJAGSINGH, S., AGOULNIK, I. U., ZHANG, Z. & LIU, Y. 2014. Base excision repair of chemotherapeutically-induced alkylated DNA damage predominantly causes contractions of expanded GAA repeats associated with Friedreich's ataxia. *PLoS ONE [Electronic Resource],* 9**,** e93464.

LAI, Y., XU, M., ZHANG, Z. & LIU, Y. 2013. Instability of CTG repeats is governed by the position of a DNA base lesion through base excision repair. *PLoS ONE [Electronic Resource],* 8**,** e56960.

LANDLES, C., SATHASIVAM, K., WEISS, A., WOODMAN, B., MOFFITT, H., FINKBEINER, S., SUN, B., GAFNI, J., ELLERBY, L. M., TROTTIER, Y., RICHARDS, W. G., OSMAND, A., PAGANETTI, P. & BATES, G. P. 2010. Proteolysis of Mutant Huntingtin Produces an Exon 1 Fragment That Accumulates as an Aggregated Protein in Neuronal Nuclei in Huntington Disease. *The Journal of Biological Chemistry,* 285**,** 8808-8823.

LANGBEHN, D. R., BRINKMAN, R. R., FALUSH, D., PAULSEN, J. S., HAYDEN, M. R. & INTERNATIONAL HUNTINGTON'S DISEASE COLLABORATIVE, G. 2004. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clin Genet,* 65**,** 267-77.

LANGBEHN, D. R., HAYDEN, M. R. & PAULSEN, J. S. 2010. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. *Am J Med Genet B Neuropsychiatr Genet,* 153b**,** 397-408.

LANGE, H., THORNER, G., HOPF, A. & SCHRODER, K. F. 1976. Morphometric studies of the neuropathological changes in choreatic diseases. *J Neurol Sci,* 28**,** 401-25.

LANGFELDER, P. & HORVATH, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics,* 9**,** 559.

LANSKA, D. J., LAVINE, L., LANSKA, M. J. & SCHOENBERG, B. S. 1988. Huntington's disease mortality in the United States. *Neurology,* 38**,** 769-72.

LAU, P. J. & KOLODNER, R. D. 2003. Transfer of the MSH2.MSH6 complex from proliferating cell nuclear antigen to mispaired bases in DNA. *J Biol Chem,* 278**,** 14-7.

LE BER, I., CAMUZAT, A., CASTELNOVO, G., AZULAY, J. P., GENTON, P., GASTAUT, J. L., BROGLIN, D., LABAUGE, P., BRICE, A. & DURR, A. 2003. Prevalence of dentatorubral-pallidoluysian atrophy in a large series of white patients with cerebellar ataxia. *Arch Neurol,* 60**,** 1097-9.

LEE, J.-M., CORREIA, K., LOUPE, J., KIM, K.-H., BARKER, D., HONG, E. P., CHAO, M. J., LONG, J. D., LUCENTE, D., VONSATTEL, J.-P., MOURO PINTO, R., ABU ELNEEL, K., RAMOS, E. M., MYSORE, J. S., GILLIS, T., WHEELER, V. C., MACDONALD, M. E., GUSELLA, J. F., MASSEY, T., MCALLISTER, B., MEDWAY, C., STONE, T. C., HALL, L., JONES, L., HOLMANS, P., KWAK, S., EHRHARDT, A., SAMPAIO, C., CIOSI, M., MAXWELL, A., CHATZI, A., MONCKTON, D. G., ORTH, M., LANDWEHRMEYER, G. B., PAULSEN, J. S., DORSEY, E. R., SHOULSON, I. & MYERS, R. H. 2019. Huntington's disease onset is determined by length of uninterrupted CAG, not encoded polyglutamine, and is modified by DNA maintenance mechanisms. 529768.

LEE, J.-M., GILLIS, T., MYSORE, JAYALAKSHMI S., RAMOS, ELIANA M., MYERS, RICHARD H., HAYDEN, MICHAEL R., MORRISON, PATRICK J., NANCE, M., ROSS, CHRISTOPHER A., MARGOLIS, RUSSELL L., SQUITIERI, F., GRIGUOLI, A., DI DONATO, S., GOMEZ-TORTOSA, E., AYUSO, C., SUCHOWERSKY, O., TRENT, RONALD J., MCCUSKER, E., NOVELLETTO, A., FRONTALI, M., JONES, R., ASHIZAWA, T., FRANK, S., SAINT-HILAIRE, M.-H., HERSCH, STEVEN M., ROSAS, HERMINIA D., LUCENTE, D., HARRISON, MADALINE B., ZANKO, A., ABRAMSON, RUTH K., MARDER, K., SEQUEIROS, J., MACDONALD, MARCY E. & GUSELLA, JAMES F. 2012a. Common SNP-Based Haplotype Analysis of the 4p16.3 Huntington Disease Gene Region. *The American Journal of Human Genetics,* 90**,** 434-444.

LEE, J. H., LEE, J. M., RAMOS, E. M., GILLIS, T., MYSORE, J. S., KISHIKAWA, S., HADZI, T., HENDRICKS, A. E., HAYDEN, M. R., MORRISON, P. J., NANCE, M., ROSS, C. A., MARGOLIS, R. L., SQUITIERI, F., GELLERA, C., GOMEZ-TORTOSA, E., AYUSO, C., SUCHOWERSKY, O., TRENT, R. J., MCCUSKER, E., NOVELLETTO, A., FRONTALI, M., JONES, R., ASHIZAWA, T., FRANK, S., SAINT-HILAIRE, M. H., HERSCH, S. M., ROSAS, H. D., LUCENTE, D., HARRISON, M. B., ZANKO, A., ABRAMSON, R. K., MARDER, K., SEQUEIROS, J., LANDWEHRMEYER, G. B., REGISTRY STUDY OF THE EUROPEAN HUNTINGTON'S DISEASE, N., SHOULSON, I., HUNTINGTON STUDY GROUP, C. P., MYERS, R. H., MACDONALD, M. E. & GUSELLA, J. F. 2012b. TAA repeat variation in the GRIK2 gene does not

influence age at onset in Huntington's disease. *Biochemical & Biophysical Research Communications,* 424**,** 404-8.

LEE, J. H., LEE, J. M., RAMOS, E. M., GILLIS, T., MYSORE, J. S., KISHIKAWA, S., HADZI, T., HENDRICKS, A. E., HAYDEN, M. R., MORRISON, P. J., NANCE, M., ROSS, C. A., MARGOLIS, R. L., SQUITIERI, F., GELLERA, C., GOMEZ-TORTOSA, E., AYUSO, C., SUCHOWERSKY, O., TRENT, R. J., MCCUSKER, E., NOVELLETTO, A., FRONTALI, M., JONES, R., ASHIZAWA, T., FRANK, S., SAINT-HILAIRE, M. H., HERSCH, S. M., ROSAS, H. D., LUCENTE, D., HARRISON, M. B., ZANKO, A., ABRAMSON, R. K., MARDER, K., SEQUEIROS, J., LANDWEHRMEYER, G. B., SHOULSON, I., MYERS, R. H., MACDONALD, M. E. & GUSELLA, J. F. 2012c. TAA repeat variation in the GRIK2 gene does not influence age at onset in Huntington's disease. *Biochem Biophys Res Commun,* 424**,** 404-8.

LEE, J. M., CHAO, M. J., HAROLD, D., ABU ELNEEL, K., GILLIS, T., HOLMANS, P., JONES, L., ORTH, M., MYERS, R. H., KWAK, S., WHEELER, V. C., MACDONALD, M. E. & GUSELLA, J. F. 2017. A modifier of Huntington's disease onset at the MLH1 locus. *Hum Mol Genet,* 26**,** 3859-3867.

LEE, J. M., PINTO, R. M., GILLIS, T., ST CLAIRE, J. C. & WHEELER, V. C. 2011a. Quantification of age-dependent somatic CAG repeat instability in Hdh CAG knock-in mice reveals different expansion dynamics in striatum and liver. *PLoS One,* 6**,** e23647.

LEE, J. M., RAMOS, E. M., LEE, J. H., GILLIS, T., MYSORE, J. S., HAYDEN, M. R., WARBY, S. C., MORRISON, P., NANCE, M., ROSS, C. A., MARGOLIS, R. L., SQUITIERI, F., OROBELLO, S., DI DONATO, S., GOMEZ-TORTOSA, E., AYUSO, C., SUCHOWERSKY, O., TRENT, R. J., MCCUSKER, E., NOVELLETTO, A., FRONTALI, M., JONES, R., ASHIZAWA, T., FRANK, S., SAINT-HILAIRE, M. H., HERSCH, S. M., ROSAS, H. D., LUCENTE, D., HARRISON, M. B., ZANKO, A., ABRAMSON, R. K., MARDER, K., SEQUEIROS, J., PAULSEN, J. S., LANDWEHRMEYER, G. B., MYERS, R. H., MACDONALD, M. E. & GUSELLA, J. F. 2012d. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology,* 78**,** 690-5.

LEE, J. M., ZHANG, J., SU, A. I., WALKER, J. R., WILTSHIRE, T., KANG, K., DRAGILEVA, E., GILLIS, T., LOPEZ, E. T., BOILY, M. J., CYR, M., KOHANE, I., GUSELLA, J. F., MACDONALD, M. E. & WHEELER, V. C. 2010. A novel approach to investigate tissue-specific trinucleotide repeat instability. *BMC Syst Biol,* 4**,** 29.

LEE, S. T., CHU, K., IM, W. S., YOON, H. J., IM, J. Y., PARK, J. E., PARK, K. H., JUNG, K. H., LEE, S. K., KIM, M. & ROH, J. K. 2011b. Altered microRNA regulation in Huntington's disease models. *Exp Neurol,* 227**,** 172-9.

LEEFLANG, E. P., ZHANG, L., TAVARE, S., HUBERT, R., SRINIDHI, J., MACDONALD, M. E., MYERS, R. H., DE YOUNG, M., WEXLER, N. S., GUSELLA, J. F. & ET AL. 1995. Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum. *Hum Mol Genet,* 4**,** 1519-26.

LEEK, J. T. 2014. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res,* 42.

LI, J. L., HAYDEN, M. R., WARBY, S. C., DURR, A., MORRISON, P. J., NANCE, M., ROSS, C. A., MARGOLIS, R. L., ROSENBLATT, A., SQUITIERI, F., FRATI, L., GOMEZ-TORTOSA, E., GARCIA, C. A., SUCHOWERSKY, O., KLIMEK, M. L., TRENT, R. J., MCCUSKER, E., NOVELLETTO, A., FRONTALI, M., PAULSEN, J. S., JONES, R., ASHIZAWA, T., LAZZARINI, A., WHEELER, V. C., PRAKASH, R., XU, G., DJOUSSE, L., MYSORE, J. S., GILLIS, T., HAKKY, M., CUPPLES, L. A., SAINT-HILAIRE, M. H., CHA, J. H., HERSCH, S. M., PENNEY, J. B., HARRISON, M. B., PERLMAN, S. L., ZANKO, A., ABRAMSON, R. K., LECHICH, A. J., DUCKETT, A., MARDER, K., CONNEALLY, P. M., GUSELLA, J. F., MACDONALD, M. E. & MYERS, R. H. 2006. Genome-wide significance for a modifier of age at neurological onset in Huntington's disease at 6q23-24: the HD MAPS study. *BMC Med Genet,* 7**,** 71.

LI, S. H., CHENG, A. L., ZHOU, H., LAM, S., RAO, M., LI, H. & LI, X. J. 2002. Interaction of Huntington disease protein with transcriptional activator Sp1. *Mol Cell Biol,* 22**,** 1277-87.

LIA, A. S., SEZNEC, H., HOFMANN-RADVANYI, H., RADVANYI, F., DUROS, C., SAQUET, C., BLANCHE, M., JUNIEN, C. & GOURDON, G. 1998. Somatic instability of the CTG repeat in mice transgenic for the myotonic dystrophy region is age dependent but not correlated to the relative intertissue transcription levels and proliferative capacities. *Hum Mol Genet,* 7**,** 1285-91.

LIMBERIS, M. P. & WILSON, J. M. 2006. Adeno-associated virus serotype 9 vectors transduce murine alveolar and nasal epithelia and can be readministered. *Proc Natl Acad Sci U S A,* 103**,** 12993-8.

LIN, J., WU, P. H., TARR, P. T., LINDENBERG, K. S., ST-PIERRE, J., ZHANG, C. Y., MOOTHA, V. K., JAGER, S., VIANNA, C. R., REZNICK, R. M., CUI, L., MANIERI, M., DONOVAN, M. X., WU, Z., COOPER, M. P., FAN, M. C., ROHAS, L. M., ZAVACKI, A. M., CINTI, S., SHULMAN, G. I., LOWELL, B. B., KRAINC, D. & SPIEGELMAN, B. M. 2004. Defects in adaptive energy metabolism with CNS-linked hyperactivity in PGC-1alpha null mice. *Cell,* 119**,** 121-35.

LIN, L., PARK, J. W., RAMACHANDRAN, S., ZHANG, Y., TSENG, Y. T., SHEN, S., WALDVOGEL, H. J., CURTIS, M. A., FAULL, R. L., TRONCOSO, J. C., PLETNIKOVA, O., ROSS, C. A., DAVIDSON, B. L. & XING, Y. 2016. Transcriptome sequencing reveals aberrant alternative splicing in Huntington's disease. *Hum Mol Genet,* 25**,** 3454-3466.

LIN, X. & ASHIZAWA, T. 2003. SCA10 and ATTCT repeat expansion: clinical features and molecular aspects. *Cytogenet Genome Res,* 100**,** 184-8.

LIU, G., BISSLER, J. J., SINDEN, R. R. & LEFFAK, M. 2007. Unstable spinocerebellar ataxia type 10 (ATTCT (AGAAT) repeats are associated with aberrant replication at the ATX10 locus and replication origin-dependent expansion at an ectopic site in human cells. *Molecular & Cellular Biology,* 27**,** 7828-38.

LIU, G., CHEN, X., BISSLER, J. J., SINDEN, R. R. & LEFFAK, M. 2010a. Replication-dependent instability at (CTG) x (CAG) repeat hairpins in human cells. *Nat Chem Biol,* 6**,** 652-9.

LIU, J., BANG, A. G., KINTNER, C., ORTH, A. P., CHANDA, S. K., DING, S. & SCHULTZ, P. G. 2005. Identification of the Wnt signaling activator leucine-rich repeat in Flightless interaction protein 2 by a genome-wide functional analysis. *Proc Natl Acad Sci U S A,* 102**,** 1927-32.

LIU, T., GHOSAL, G., YUAN, J., CHEN, J. & HUANG, J. 2010b. FAN1 acts with FANCI-FANCD2 to promote DNA interstrand cross-link repair. *Science,* 329**,** 693-6.

LIU, T., GHOSAL, G., YUAN, J., CHEN, J. & HUANG, J. 2010c. FAN1 acts with FANCI-FANCD2 to promote DNA interstrand cross-link repair. *Science.* United States.

LIU, Y. & WILSON, S. H. 2012. DNA base excision repair: a mechanism of trinucleotide repeat expansion. *Trends Biochem Sci,* 37**,** 162-72.

LOKANGA, R. A., ZHAO, X. N. & USDIN, K. 2014. The mismatch repair protein MSH2 is rate limiting for repeat expansion in a fragile X premutation mouse model. *Hum Mutat,* 35**,** 129-36.

LOPES-CENDES, I., MACIEL, P., KISH, S., GASPAR, C., ROBITAILLE, Y., CLARK, H. B., KOEPPEN, A. H., NANCE, M., SCHUT, L., SILVEIRA, I., COUTINHO, P., SEQUEIROS, J. & ROULEAU, G. A. 1996. Somatic mosaicism in the central nervous system in spinocerebellar ataxia type 1 and Machado-Joseph disease. *Ann Neurol,* 40**,** 199-206.

LOPEZ CASTEL, A., CLEARY, J. D. & PEARSON, C. E. 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol,* 11**,** 165-70.

LOPEZ CASTEL, A., NAKAMORI, M., TOME, S., CHITAYAT, D., GOURDON, G., THORNTON, C. A. & PEARSON, C. E. 2011. Expanded CTG repeat demarcates a boundary for abnormal CpG methylation in myotonic dystrophy patient tissues. *Hum Mol Genet,* 20**,** 1-15.

LOPEZ CASTEL, A., TOMKINSON, A. E. & PEARSON, C. E. 2009. CTG/CAG repeat instability is modulated by the levels of human DNA ligase I and its interaction with proliferating cell nuclear antigen: a distinction between replication and slipped-DNA repair. *J Biol Chem,* 284**,** 26631-45.

LORENZETTI, D., BOHLEGA, S. & ZOGHBI, H. Y. 1997. The expansion of the CAG repeat in ataxin-2 is a frequent cause of autosomal dominant spinocerebellar ataxia. *Neurology,* 49**,** 1009-13.

LOSEKOOT, M., VAN BELZEN, M. J., SENECA, S., BAUER, P., STENHOUSE, S. A. & BARTON, D. E. 2013. EMQN/CMGS best practice guidelines for the molecular genetic testing of Huntington disease. *Eur J Hum Genet,* 21**,** 480-6.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15**,** 550.

LOVRECIC, L., KASTRIN, A., KOBAL, J., PIRTOSEK, Z., KRAINC, D. & PETERLIN, B. 2009. Gene expression changes in blood as a putative biomarker for Huntington's disease. *Mov Disord,* 24**,** 2277-81.

LU, T., PAN, Y., KAO, S. Y., LI, C., KOHANE, I., CHAN, J. & YANKNER, B. A. 2004. Gene regulation and DNA damage in the ageing human brain. *Nature,* 429**,** 883-91.

LUTHI-CARTER, R., STRAND, A. D., HANSON, S. A., KOOPERBERG, C., SCHILLING, G., LA SPADA, A. R., MERRY, D. E., YOUNG, A. B., ROSS, C. A., BORCHELT, D. R. & OLSON, J. M. 2002. Polyglutamine and transcription: gene expression changes shared by DRPLA and Huntington's disease mouse models reveal context-independent effects. *Hum Mol Genet,* 11**,** 1927-37.

LYNDAKER, A. M. & ALANI, E. 2009. A tale of tails: insights into the coordination of 3′ end processing during homologous recombination. *BioEssays,* 31**,** 315-321.

MACDONALD, M. E., VONSATTEL, J. P., SHRINIDHI, J., COUROPMITREE, N. N., CUPPLES, L. A., BIRD, E. D., GUSELLA, J. F. & MYERS, R. H. 1999. Evidence for the GluR6 gene associated with younger onset age of Huntington's disease. *Neurology,* 53**,** 1330-2.

MACKAY, C., DECLAIS, A. C., LUNDIN, C., AGOSTINHO, A., DEANS, A. J., MACARTNEY, T. J., HOFMANN, K., GARTNER, A., WEST, S. C., HELLEDAY, T., LILLEY, D. M. & ROUSE, J. 2010a. Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. *Cell.* United States: 2010 Elsevier Inc.

MACKAY, C., DECLAIS, A. C., LUNDIN, C., AGOSTINHO, A., DEANS, A. J., MACARTNEY, T. J., HOFMANN, K., GARTNER, A., WEST, S. C., HELLEDAY, T., LILLEY, D. M. & ROUSE, J. 2010b. Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. *Cell,* 142**,** 65-76.

MADABHUSHI, R., GAO, F., PFENNING, A. R., PAN, L., YAMAKAWA, S., SEO, J., RUEDA, R., PHAN, T. X., YAMAKAWA, H., PAO, P. C., STOTT, R. T., GJONESKA, E., NOTT, A., CHO, S., KELLIS, M. & TSAI, L. H. 2015. Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. *Cell,* 161**,** 1592-605.

MADABHUSHI, R., PAN, L. & TSAI, L. H. 2014. DNA damage and its links to neurodegeneration. *Neuron,* 83**,** 266-282.

MANGIARINI, L., SATHASIVAM, K., MAHAL, A., MOTT, R., SELLER, M. & BATES, G. P. 1997. Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. *Nat Genet,* 15**,** 197-200.

MANGIARINI, L., SATHASIVAM, K., SELLER, M., COZENS, B., HARPER, A., HETHERINGTON, C., LAWTON, M., TROTTIER, Y., LEHRACH, H., DAVIES, S. W. & BATES, G. P. 1996. Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell,* 87**,** 493-506.

MANLEY, K., SHIRLEY, T. L., FLAHERTY, L. & MESSER, A. 1999. Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat Genet,* 23**,** 471-3.

MARGOLIS, R. L., STINE, O. C., CALLAHAN, C., ROSENBLATT, A., ABBOTT, M. H., SHERR, M. & ROSS, C. A. 1999. Two novel single-base-pair substitutions adjacent to the CAG repeat in the huntington disease gene (IT15): implications for diagnostic testing. *Am J Hum Genet,* 64**,** 323-6.

MARIANI, L. L., TESSON, C., CHARLES, P., CAZENEUVE, C., HAHN, V., YOUSSOV, K., FREEMAN, L., GRABLI, D., ROZE, E., NOEL, S., PEUVION, J. N., BACHOUD-LEVI, A. C., BRICE, A., STEVANIN, G. & DURR, A. 2016. Expanding the Spectrum of Genes Involved in Huntington Disease Using a Combined Clinical and Genetic Approach. *JAMA Neurol,* 73**,** 1105-14.

MARRA, G., IACCARINO, I., LETTIERI, T., ROSCILLI, G., DELMASTRO, P. & JIRICNY, J. 1998. Mismatch repair deficiency associated with overexpression of the MSH3 gene. *Proc Natl Acad Sci U S A,* 95**,** 8568-73.

MARTI, E., PANTANO, L., BANEZ-CORONEL, M., LLORENS, F., MINONES-MOYANO, E., PORTA, S., SUMOY, L., FERRER, I. & ESTIVILL, X. 2010. A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res,* 38**,** 7219-35.

MARTINS, S., PEARSON, C. E., COUTINHO, P., PROVOST, S., AMORIM, A., DUBE, M. P., SEQUEIROS, J. & ROULEAU, G. A. 2014. Modifiers of (CAG)(n) instability in Machado-Joseph disease (MJD/SCA3) transmissions: an association study with DNA replication, repair and recombination genes. *Hum Genet,* 133**,** 1311-8.

MARTORELL, L., MARTINEZ, J. M., CAREY, N., JOHNSON, K. & BAIGET, M. 1995. Comparison of CTG repeat length expansion and clinical progression of myotonic dystrophy over a five year period. *J Med Genet,* 32**,** 593-6.

MARUFF, P., TYLER, P., BURT, T., CURRIE, B., BURNS, C. & CURRIE, J. 1996. Cognitive deficits in Machado-Joseph disease. *Ann Neurol,* 40**,** 421-7.

MARUYAMA, H., NAKAMURA, S., MATSUYAMA, Z., SAKAI, T., DOYU, M., SOBUE, G., SETO, M., TSUJIHATA, M., OH-I, T., NISHIO, T. & ET AL. 1995. Molecular features of the CAG repeats and clinical manifestation of Machado-Joseph disease. *Hum Mol Genet,* 4**,** 807-12.

MASSEY, T. H. & JONES, L. 2018. The central role of DNA damage and repair in CAG repeat diseases. *Dis Model Mech,* 11.

MASTROKOLIAS, A., ARIYUREK, Y., GOEMAN, J. J., VAN DUIJN, E., ROOS, R. A., VAN DER MAST, R. C., VAN OMMEN, G. B., DEN DUNNEN, J. T., T HOEN, P. A. & VAN ROON-MOM, W. M. 2015. Huntington's disease biomarker progression profile identified by transcriptome sequencing in peripheral blood. *Eur J Hum Genet*.

MATHESON, E. C., HOGARTH, L. A., CASE, M. C., IRVING, J. A. & HALL, A. G. 2007. DHFR and MSH3 co-amplification in childhood acute lymphoblastic leukaemia, in vitro and in vivo. *Carcinogenesis,* 28**,** 1341-6.

MATSUYAMA, Z., KAWAKAMI, H., MARUYAMA, H., IZUMI, Y., KOMURE, O., UDAKA, F., KAMEYAMA, M., NISHIO, T., KURODA, Y., NISHIMURA, M. & NAKAMURA, S. 1997. Molecular features of the CAG repeats of spinocerebellar ataxia 6 (SCA6). *Hum Mol Genet,* 6**,** 1283-7.

MATTIS, V. B., TOM, C., AKIMOV, S., SAEEDIAN, J., OSTERGAARD, M. E., SOUTHWELL, A. L., DOTY, C. N., ORNELAS, L., SAHABIAN, A., LENAEUS, L., MANDEFRO, B., SAREEN, D., ARJOMAND, J., HAYDEN, M. R., ROSS, C. A. & SVENDSEN, C. N. 2015. HD iPSC-derived neural progenitors accumulate in culture and are susceptible to BDNF withdrawal due to glutamate toxicity. *Hum Mol Genet,* 24**,** 3257-71.

MAUCKSCH, C., VAZEY, E. M., GORDON, R. J. & CONNOR, B. 2013. Stem cell-based therapy for Huntington's disease. *J Cell Biochem,* 114**,** 754-63.

MCCABE, K. M., OLSON, S. B. & MOSES, R. E. 2009. DNA interstrand crosslink repair in mammalian cells. *J Cell Physiol,* 220**,** 569-73.

MCCAMPBELL, A., TAYLOR, J. P., TAYE, A. A., ROBITSCHEK, J., LI, M., WALCOTT, J., MERRY, D., CHAI, Y., PAULSON, H., SOBUE, G. & FISCHBECK, K. H. 2000. CREB-binding protein sequestration by expanded polyglutamine. *Hum Mol Genet,* 9**,** 2197-202.

MCCOLGAN, P., SEUNARINE, K. K., RAZI, A., COLE, J. H., GREGORY, S., DURR, A., ROOS, R. A., STOUT, J. C., LANDWEHRMEYER, B., SCAHILL, R. I., CLARK, C. A., REES, G. & TABRIZI, S. J. 2015. Selective vulnerability of Rich Club brain regions is an organizational principle of structural connectivity loss in Huntington's disease. *Brain,* 138**,** 3327-44.

MCCONNELL, M. J., LINDBERG, M. R., BRENNAND, K. J., PIPER, J. C., VOET, T., COWING-ZITRON, C., SHUMILINA, S., LASKEN, R. S., VERMEESCH, J. R., HALL, I. M. & GAGE, F. H. 2013. Mosaic copy number variation in human neurons. *Science,* 342**,** 632-7.

MCKINNON, P. J. 2009. DNA repair deficiency and neurological disease. *Nat Rev Neurosci,* 10**,** 100-12.

MCKUSICK, V. A. 2007. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet,* 80**,** 588-604.

MCMURRAY, C. T. 2008. Hijacking of the mismatch repair system to cause CAG expansion and cell death in neurodegenerative disease. *DNA Repair (Amst),* 7**,** 1121-34.

MCMURRAY, C. T. 2010. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet,* 11**,** 786-99.

MEETEI, A. R., MEDHURST, A. L., LING, C., XUE, Y., SINGH, T. R., BIER, P., STELTENPOOL, J., STONE, S., DOKAL, I., MATHEW, C. G., HOATLIN, M., JOENJE, H., DE WINTER, J. P. & WANG, W. 2005. A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. *Nat Genet,* 37**,** 958-63.

MENON, R. P., NETHISINGHE, S., FAGGIANO, S., VANNOCCI, T., REZAEI, H., PEMBLE, S., SWEENEY, M. G., WOOD, N. W., DAVIS, M. B., PASTORE, A. & GIUNTI, P. 2013. The role of interruptions in polyQ in the pathology of SCA1. *PLoS Genet,* 9**,** e1003648.

MEOLA, G. & CARDANI, R. 2015. Myotonic dystrophies: An update on clinical aspects, genetic, pathology, and molecular pathomechanisms. *Biochim Biophys Acta,* 1852**,** 594-606.

METZGER, S., RONG, J., NGUYEN, H. P., CAPE, A., TOMIUK, J., SOEHN, A. S., PROPPING, P., FREUDENBERG-HUA, Y., FREUDENBERG, J., TONG, L., LI, S. H., LI, X. J. & RIESS, O. 2008. Huntingtin-associated protein-1 is a modifier of the age-at-onset of Huntington's disease. *Hum Mol Genet,* 17**,** 1137-46.

METZGER, S., SAUKKO, M., VAN CHE, H., TONG, L., PUDER, Y., RIESS, O. & NGUYEN, H. P. 2010. Age at onset in Huntington's disease is modified by the autophagy pathway: implication of the V471A polymorphism in Atg7. *Hum Genet,* 128**,** 453-9.

MGI. 2016. *MGI-Mouse Genome Informatics-The international database resource for the laboratory mouse* [Online]. Available: http://www.informatics.jax.org/ [Accessed 21/01/2016].

MIELCAREK, M., LANDLES, C., WEISS, A., BRADAIA, A., SEREDENINA, T., INUABASI, L., OSBORNE, G. F., WADEL, K., TOULLER, C., BUTLER, R., ROBERTSON, J., FRANKLIN, S. A., SMITH, D. L., PARK, L., MARKS, P. A., WANKER, E. E., OLSON, E. N., LUTHI-CARTER, R., VAN DER PUTTEN, H., BEAUMONT, V. & BATES, G. P. 2013. HDAC4 reduction: a novel therapeutic strategy to target cytoplasmic huntingtin and ameliorate neurodegeneration. *PLoS Biol,* 11**,** e1001717.

MIHM, M. J., AMANN, D. M., SCHANBACHER, B. L., ALTSCHULD, R. A., BAUER, J. A. & HOYT, K. R. 2007. Cardiac dysfunction in the R6/2 mouse model of Huntington's disease. *Neurobiology of Disease,* 25**,** 297-308.

MILLER, B. R., DORNER, J. L., SHOU, M., SARI, Y., BARTON, S. J., SENGELAUB, D. R., KENNEDY, R. T. & REBEC, G. V. 2008. Up-regulation of GLT1 expression increases glutamate uptake and attenuates the Huntington's disease phenotype in the R6/2 mouse. *Neuroscience,* 153**,** 329-37.

MILLER, J. R., LO, K. K., ANDRE, R., HENSMAN MOSS, D. J., TRAGER, U., STONE, T. C., JONES, L., HOLMANS, P., PLAGNOL, V. & TABRIZI, S. J. 2016. RNA-Seq of Huntington's disease patient myeloid cells reveals innate transcriptional dysregulation associated with proinflammatory pathway activation. *Hum Mol Genet*.

MILLIPORE, M. 2016. *ReNcell VM Human Neural Progenitor Cell Line | SCC008* [Online]. Human ventral mesencephalon brain tissue. Available: https://www.merckmillipore.com/GB/en/product/ReNcell-VM-Human-Neural-Progenitor-Cell-LIne,MM_NF-SCC008?ReferrerURL=https%3A%2F%2Fwww.google.co.uk%2F&bd=1 [Accessed 19/12/2016].

MILNE, I., STEPHEN, G., BAYER, M., COCK, P. J., PRITCHARD, L., CARDLE, L., SHAW, P. D. & MARSHALL, D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform,* 14**,** 193-202.

MINA, E., VAN ROON-MOM, W. M., HETTNE, K. M., VAN ZWET, E., GOEMAN, J. J., NERI, C., MONS, B., T'HOEN, P. A. C. & ROOS, M. 2016. Common disease signatures between blood and brain in Huntington's Disease. *Orphanet Journal of Rare Diseases*.

MIRKIN, S. M. 2007. Expandable DNA repeats and human disease. *Nature,* 447**,** 932-40.

MOCHEL, F. & HALLER, R. G. 2011. Energy deficit in Huntington disease: why it matters. *J Clin Invest,* 121**,** 493-9.

MOLLERSEN, L., ROWE, A. D., LARSEN, E., ROGNES, T. & KLUNGLAND, A. 2010. Continuous and periodic expansion of CAG repeats in Huntington's disease R6/1 mice. *PLoS Genet,* 6**,** e1001242.

MOLLICA, P. A., REID, J. A., OGLE, R. C., SACHS, P. C. & BRUNO, R. D. 2016. DNA Methylation Leads to DNA Repair Gene Down-Regulation and Trinucleotide Repeat Expansion in Patient-Derived Huntington Disease Cells. *Am J Pathol,* 186**,** 1967-1976.

MONTANINI, L., FERRARI, S., CRAFA, P., GHIRARDINI, S., PONZIN, D., ORSONI, J. G. & MORA, P. 2013. Human RNA integrity after postmortem retinal tissue recovery. *Ophthalmic Genet,* 34**,** 27-31.

MORALES, F., COUTO, J. M., HIGHAM, C. F., HOGG, G., CUENCA, P., BRAIDA, C., WILSON, R. H., ADAM, B., DEL VALLE, G., BRIAN, R., SITTENFELD, M., ASHIZAWA, T., WILCOX, A., WILCOX, D. E. & MONCKTON, D. G. 2012. Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Hum Mol Genet,* 21**,** 3558-67.

MORALES, F., VASQUEZ, M., SANTAMARIA, C., CUENCA, P., CORRALES, E. & MONCKTON, D. G. 2016. A polymorphism in the MSH3 mismatch repair gene is associated with the levels of somatic instability of the expanded CTG repeat in the blood DNA of myotonic dystrophy type 1 patients. *DNA Repair (Amst),* 40**,** 57-66.

MORALES, F. A. 2006. *Somatic mosaicism and genotype-phenotype correlations in myotonic dystrophy type 1.* Ph.D., University of Glasgow.

MORRISON, P. J. 2012. Prevalence estimates of Huntington disease in Caucasian populations are gross underestimates. *Mov Disord,* 27**,** 1707-8; author reply 1708-9.

MORTON, A. J. 2013. Circadian and sleep disorder in Huntington's disease. *Exp Neurol,* 243**,** 34-44.

MORTON, A. J. & HOWLAND, D. S. 2013. Large genetic animal models of Huntington's Disease. *J Huntingtons Dis,* 2**,** 3-19.

MRZLJAK, L. & MUNOZ-SANJUAN, I. 2015. Therapeutic Strategies for Huntington's Disease. *Curr Top Behav Neurosci,* 22**,** 161-201.

MUNOZ, I. M., SZYNIAROWSKI, P., TOTH, R., ROUSE, J. & LACHAUD, C. 2014. Improved genome editing in human cell lines using the CRISPR method. *PLoS One,* 9**,** e109752.

MURO, Y., SUGIURA, K., MIMORI, T. & AKIYAMA, M. 2015. DNA mismatch repair enzymes: genetic defects and autoimmunity. *Clin Chim Acta,* 442**,** 102-9.

MYERS, R. H., SAX, D. S., KOROSHETZ, W. J., MASTROMAURO, C., CUPPLES, L. A., KIELY, D. K., PETTENGILL, F. K. & BIRD, E. D. 1991. Factors associated with slow progression in Huntington's disease. *Arch Neurol,* 48**,** 800-4.

NAGEL, Z. D., MARGULIES, C. M., CHAIM, I. A., MCREE, S. K., MAZZUCATO, P., AHMAD, A., ABO, R. P., BUTTY, V. L., FORGET, A. L. & SAMSON, L. D. 2014. Multiplexed DNA repair assays for multiple lesions and multiple doses via transcription inhibition and transcriptional mutagenesis. *Proc Natl Acad Sci U S A,* 111**,** E1823-32.

NAITO, H. & OYANAGI, S. 1982. Familial myoclonus epilepsy and choreoathetosis: hereditary dentatorubral-pallidoluysian atrophy. *Neurology,* 32**,** 798-807.

NAKAJIMA, E., ORIMO, H., IKEJIMA, M. & SHIMADA, T. 1995. Nine-bp repeat polymorphism in exon 1 of the hMSH3 gene. *Jpn J Hum Genet,* 40**,** 343-5.

NAKAMORI, M., PEARSON, C. E. & THORNTON, C. A. 2011. Bidirectional transcription stimulates expansion and contraction of expanded (CTG) (CAG) repeats. *Human Molecular Genetics,* 20**,** 580-8.

NAKATANI, R., NAKAMORI, M., FUJIMURA, H., MOCHIZUKI, H. & TAKAHASHI, M. P. 2015a. Large expansion of CTG CAG repeats is exacerbated by MutSbeta in human cells. *Scientific Reports,* 5**,** 11020.

NAKATANI, R., NAKAMORI, M., FUJIMURA, H., MOCHIZUKI, H. & TAKAHASHI, M. P. 2015b. Large expansion of CTG*CAG repeats is exacerbated by MutSbeta in human cells. *Sci Rep,* 5**,** 11020.

NAKATANI, R., NAKAMORI, M., FUJIMURA, H., MOCHIZUKI, H. & TAKAHASHI, M. P. 2015c. Large expansion of CTG•CAG repeats is exacerbated by MutSβ in human cells. *Sci Rep,* 5.

NANCE, M. A., MATHIAS-HAGEN, V., BRENINGSTALL, G., WICK, M. J. & MCGLENNEN, R. C. 1999. Analysis of a very large trinucleotide repeat in a patient with juvenile Huntington's disease. *Neurology,* 52**,** 392-4.

NAZE, P., VUILLAUME, I., DESTEE, A., PASQUIER, F. & SABLONNIERE, B. 2002. Mutation analysis and association studies of the ubiquitin carboxy-terminal hydrolase L1 gene in Huntington's disease. *Neurosci Lett,* 328**,** 1-4.

NEIL, A. J., KIM, J. C. & MIRKIN, S. M. 2017. Precarious maintenance of simple DNA repeats in eukaryotes. *Bioessays,* 39.

NENGUKE, T., ALADJEM, M. I., GUSELLA, J. F., WEXLER, N. S. & ARNHEIM, N. 2003. Candidate DNA replication initiation regions at human trinucleotide repeat disease loci. *Hum Mol Genet,* 12**,** 1021-8.

NEUEDER, A. & BATES, G. P. 2014. A common gene expression signature in Huntington's disease patient brain regions. *BMC Med Genomics,* 7**,** 60.

NIEDERNHOFER, L. J., ODIJK, H., BUDZOWSKA, M., VAN DRUNEN, E., MAAS, A., THEIL, A. F., DE WIT, J., JASPERS, N. G., BEVERLOO, H. B., HOEIJMAKERS, J. H. & KANAAR, R. 2004. The structure-specific endonuclease Ercc1-Xpf is required to resolve DNA interstrand cross-link-induced double-strand breaks. *Mol Cell Biol,* 24**,** 5776-87.

NITHIANANTHARAJAH, J. & HANNAN, A. J. 2013. Dysregulation of synaptic proteins, dendritic spine abnormalities and pathological plasticity of synapses as experience-dependent mediators of cognitive and psychiatric symptoms in Huntington's disease. *Neuroscience,* 251**,** 66-74.

NOLL, D. M., MASON, T. M. & MILLER, P. S. 2006. Formation and repair of interstrand cross-links in DNA. *Chem Rev,* 106**,** 277-301.

NORREMOLLE, A., BUDTZ-JORGENSEN, E., FENGER, K., NIELSEN, J. E., SORENSEN, S. A. & HASHOLT, L. 2009. 4p16.3 haplotype modifying age at onset of Huntington disease. *Clin Genet,* 75**,** 244-50.

NUCIFORA, F. C., JR., SASAKI, M., PETERS, M. F., HUANG, H., COOPER, J. K., YAMADA, M., TAKAHASHI, H., TSUJI, S., TRONCOSO, J., DAWSON, V. L., DAWSON, T. M. & ROSS, C. A. 2001. Interference by huntingtin and atrophin-1 with cbp-mediated transcription leading to cellular toxicity. *Science,* 291**,** 2423-8.

OLMOS-ALONSO, A., SCHETTERS, S. T., SRI, S., ASKEW, K., MANCUSO, R., VARGAS-CABALLERO, M., HOLSCHER, C., PERRY, V. H. & GOMEZ-NICOLA, D. 2016. Pharmacological targeting of CSF1R inhibits microglial proliferation and prevents the progression of Alzheimer's-like pathology. *Brain,* 139**,** 891-907.

OMIM. 2015. *LYNCH SYNDROME* [Online]. OMIM. Available: http://www.omim.org/entry/120435 [Accessed].

ONLINE MENDELIAN INHERITANCE IN MAN, O. 2015. *OMIM - Online Mendelian Inheritance in Man* [Online]. Available: http://www.omim.org/ [Accessed].

ORR, H. T. & ZOGHBI, H. Y. 2007. Trinucleotide repeat disorders. *Annu Rev Neurosci,* 30**,** 575-621.

ORTEGA, Z. & LUCAS, J. J. 2014. Ubiquitin-proteasome system involvement in Huntington's disease. *Front Mol Neurosci,* 7**,** 77.

ORTH, M., COOPER, J. M., BATES, G. P. & SCHAPIRA, A. H. V. 2003. Inclusion formation in Huntington's disease R6/2 mouse muscle cultures. *Journal of Neurochemistry,* 87**,** 1-6.

ORTH, M., EUROPEAN HUNTINGTON'S DISEASE, N., HANDLEY, O. J., SCHWENKE, C., DUNNETT, S., WILD, E. J., TABRIZI, S. J. & LANDWEHRMEYER, G. B. 2011. Observing Huntington's disease: the European Huntington's Disease Network's REGISTRY. *J Neurol Neurosurg Psychiatry,* 82**,** 1409-12.

ORTH, M., HANDLEY, O. J., SCHWENKE, C., DUNNETT, S. B., CRAUFURD, D., HO, A. K., WILD, E., TABRIZI, S. J. & LANDWEHRMEYER, G. B. 2010. Observing Huntington's Disease: the European Huntington's Disease Network's REGISTRY. *PLoS Curr,* 2**,** Rrn1184.

OUIMET, C. C., MILLER, P. E., HEMMINGS, H. C., JR., WALAAS, S. I. & GREENGARD, P. 1984. DARPP-32, a dopamine- and adenosine 3':5'-monophosphate-regulated phosphoprotein enriched in dopamine-innervated brain regions. III. Immunocytochemical localization. *J Neurosci,* 4**,** 111-24.

OWEN, B. A., YANG, Z., LAI, M., GAJEC, M., BADGER, J. D., 2ND, HAYES, J. J., EDELMANN, W., KUCHERLAPATI, R., WILSON, T. M. & MCMURRAY, C. T. 2005. (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. *Nat Struct Mol Biol,* 12**,** 663-70.

PACKER, A. N., XING, Y., HARPER, S. Q., JONES, L. & DAVIDSON, B. L. 2008. The bifunctional microRNA miR-9/miR-9* regulates REST and CoREST and is downregulated in Huntington's disease. *J Neurosci,* 28**,** 14341-6.

PAL, T., PERMUTH-WEY, J. & SELLERS, T. A. 2008. A review of the clinical relevance of mismatch-repair deficiency in ovarian cancer. *Cancer,* 113**,** 733-42.

PANDIT, B., ROY, M., DUTTA, J., PADHI, B. K., BHOUMIK, G. & BHATTACHARYYA, N. P. 2001. Co-amplification of dhfr and a homologue of hmsh3 in a Chinese hamster methotrexate-resistant cell line correlates with resistance to a range of chemotherapeutic drugs. *Cancer Chemother Pharmacol,* 48**,** 312-8.

PANIGRAHI, G. B., CLEARY, J. D. & PEARSON, C. E. 2002. In vitro (CTG)*(CAG) expansions and deletions by human cell extracts. *J Biol Chem,* 277**,** 13926-34.

PANTHER. 2016. *PANTHER - Gene List Analysis* [Online]. Available: http://pantherdb.org/ [Accessed 21/01/2016].

PAPATHEODOROU, I., FONSECA, N. A., KEAYS, M., TANG, Y. A., BARRERA, E., BAZANT, W., BURKE, M., FULLGRABE, A., FUENTES, A. M., GEORGE, N., HUERTA, L., KOSKINEN, S., MOHAMMED, S., GENIZA, M., PREECE, J., JAISWAL, P., JARNUCZAK, A. F., HUBER, W., STEGLE, O., VIZCAINO, J. A., BRAZMA, A. & PETRYSZAK, R. 2018. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res,* 46**,** D246-D251.

PAPOUTSI, M., LABUSCHAGNE, I., TABRIZI, S. J. & STOUT, J. C. 2014. The cognitive burden in Huntington's disease: pathology, phenotype, and mechanisms of compensation. *Mov Disord,* 29**,** 673-83.

PATTISON, J. S., SANBE, A., MALOYAN, A., OSINSKA, H., KLEVITSKY, R. & ROBBINS, J. 2008. Cardiomyocyte Expression of a Polyglutamine Preamyloid Oligomer Causes Heart Failure. *Circulation,* 117**,** 2743-2751.

PAULSEN, J. S., HAYDEN, M., STOUT, J. C., LANGBEHN, D. R., AYLWARD, E., ROSS, C. A., GUTTMAN, M., NANCE, M., KIEBURTZ, K., OAKES, D., SHOULSON, I., KAYSON, E., JOHNSON, S., PENZINER, E. & PREDICT, H. D. I. O. T. H. S. G. 2006. Preparing for preventive clinical trials: the Predict-HD study. *Arch Neurol,* 63**,** 883-90.

PAULSEN, J. S., LANGBEHN, D. R., STOUT, J. C., AYLWARD, E., ROSS, C. A., NANCE, M., GUTTMAN, M., JOHNSON, S., MACDONALD, M., BEGLINGER, L. J., DUFF, K., KAYSON, E., BIGLAN, K., SHOULSON, I., OAKES, D. & HAYDEN, M. 2008. Detection of Huntington's disease decades before diagnosis: the Predict-HD study. *J Neurol Neurosurg Psychiatry,* 79**,** 874-80.

PAULSON, H. 2018. Repeat expansion diseases. *Handb Clin Neurol,* 147**,** 105-123.

PEARL, L. H., SCHIERZ, A. C., WARD, S. E., AL-LAZIKANI, B. & PEARL, F. M. 2015. Therapeutic opportunities within the DNA damage response. *Nat Rev Cancer,* 15**,** 166-80.

PEARSON, C. E., EDAMURA, K. N. & CLEARY, J. D. 2005a. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet,* 6**,** 729-742.

PEARSON, C. E., EICHLER, E. E., LORENZETTI, D., KRAMER, S. F., ZOGHBI, H. Y., NELSON, D. L. & SINDEN, R. R. 1998. Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry,* 37**,** 2701-8.

PEARSON, C. E., EWEL, A., ACHARYA, S., FISHEL, R. A. & SINDEN, R. R. 1997. Human MSH2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases. *Hum Mol Genet,* 6**,** 1117-23.

PEARSON, C. E., NICHOL EDAMURA, K. & CLEARY, J. D. 2005b. Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet,* 6**,** 729-42.

PECHO-VRIESELING, E., RIEKER, C., FUCHS, S., BLECKMANN, D., ESPOSITO, M. S., BOTTA, P., GOLDSTEIN, C., BERNHARD, M., GALIMBERTI, I., MULLER, M., LUTHI, A., ARBER, S., BOUWMEESTER, T., VAN DER PUTTEN, H. & DI GIORGIO, F. P. 2014. Transneuronal propagation of mutant huntingtin contributes to non-cell autonomous pathology in neurons. *Nat Neurosci,* 17**,** 1064-72.

PENG, M., XIE, J., UCHER, A., STAVNEZER, J. & CANTOR, S. B. 2014. Crosstalk between BRCA-Fanconi anemia and mismatch repair pathways prevents MSH2-dependent aberrant DNA damage responses. *Embo j,* 33**,** 1698-712.

PENNELL, S., DECLAIS, A. C., LI, J., HAIRE, L. F., BERG, W., SALDANHA, J. W., TAYLOR, I. A., ROUSE, J., LILLEY, D. M. & SMERDON, S. J. 2014. FAN1 activity on asymmetric repair intermediates is mediated by an atypical monomeric virus-type replication-repair nuclease domain. *Cell Rep,* 8**,** 84-93.

PENNEY, J. B., JR., VONSATTEL, J. P., MACDONALD, M. E., GUSELLA, J. F. & MYERS, R. H. 1997. CAG repeat number governs the development rate of pathology in Huntington's disease. *Ann Neurol,* 41**,** 689-92.

PERSICHETTI, F., CARLEE, L., FABER, P. W., MCNEIL, S. M., AMBROSE, C. M., SRINIDHI, J., ANDERSON, M., BARNES, G. T., GUSELLA, J. F. & MACDONALD, M. E. 1996. Differential expression of normal and mutant Huntington's disease gene alleles. *Neurobiol Dis,* 3**,** 183-90.

PERSICHETTI, F., SRINIDHI, J., KANALEY, L., GE, P., MYERS, R. H., D'ARRIGO, K., BARNES, G. T., MACDONALD, M. E., VONSATTEL, J. P., GUSELLA, J. F. & ET AL. 1994. Huntington's disease CAG trinucleotide repeats in pathologically confirmed post-mortem brains. *Neurobiol Dis,* 1**,** 159-66.

PHAROS, H. S. G. P. I.-. 2006. At risk for Huntington disease: The PHAROS (Prospective Huntington At Risk Observational Study) cohort enrolled. *Arch Neurol,* 63**,** 991-6.

PINTO, R. M., DRAGILEVA, E., KIRBY, A., LLORET, A., LOPEZ, E., ST CLAIRE, J., PANIGRAHI, G. B., HOU, C., HOLLOWAY, K., GILLIS, T., GUIDE, J. R., COHEN, P. E., LI, G. M., PEARSON, C. E., DALY, M. J. & WHEELER, V. C. 2013a. Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. *PLoS Genet,* 9**,** e1003930.

PINTO, R. M., DRAGILEVA, E., KIRBY, A., LLORET, A., LOPEZ, E., ST. CLAIRE, J., PANIGRAHI, G. B., HOU, C., HOLLOWAY, K., GILLIS, T., GUIDE, J. R., COHEN, P. E., LI, G.-M., PEARSON, C. E., DALY, M. J. & WHEELER, V. C. 2013b. Mismatch Repair Genes Mlh1 and Mlh3 Modify CAG Instability in Huntington's Disease Mice: Genome-Wide and Candidate Approaches. *PLoS Genetics,* 9**,** e1003930.

PIZZOLATO, J., MUKHERJEE, S., SCHARER, O. D. & JIRICNY, J. 2015. FANCD2-associated nuclease 1, but not exonuclease 1 or flap endonuclease 1, is able to unhook DNA interstrand cross-links in vitro. *J Biol Chem,* 290**,** 22602-11.

POLLARD, M. O., GURDASANI, D., MENTZER, A. J., PORTER, T. & SANDHU, M. S. 2018. Long reads: their purpose and place. *Hum Mol Genet,* 27**,** R234-r241.

PONTARIN, G., FERRARO, P., BEE, L., REICHARD, P. & BIANCHI, V. 2012. Mammalian ribonucleotide reductase subunit p53R2 is required for mitochondrial DNA replication and DNA repair in quiescent cells. *Proc Natl Acad Sci U S A,* 109**,** 13302-7.

PONTARIN, G., FERRARO, P., RAMPAZZO, C., KOLLBERG, G., HOLME, E., REICHARD, P. & BIANCHI, V. 2011. Deoxyribonucleotide metabolism in cycling and resting human fibroblasts with a missense mutation in p53R2, a subunit of ribonucleotide reductase. *J Biol Chem,* 286**,** 11132-40.

PORRO, A., BERTI, M., PIZZOLATO, J., BOLOGNA, S., KADEN, S., SAXER, A., MA, Y., NAGASAWA, K., SARTORI, A. A. & JIRICNY, J. 2017. FAN1 interaction with ubiquitylated PCNA alleviates replication stress and preserves genomic integrity independently of BRCA2. *Nat Commun,* 8**,** 1073.

POTTER, N. T. 1996. The relationship between (CAG)n repeat number and age of onset in a family with dentatorubral-pallidoluysian atrophy (DRPLA): diagnostic implications of confirmatory and predictive testing. *J Med Genet,* 33**,** 168-70.

PUJANA, M. A., CORRAL, J., GRATACOS, M., COMBARROS, O., BERCIANO, J., GENIS, D., BANCHS, I., ESTIVILL, X. & VOLPINI, V. 1999. Spinocerebellar ataxias in Spanish patients: genetic analysis of familial and sporadic cases. The Ataxia Study Group. *Hum Genet,* 104**,** 516-22.

PULST, S. M., SANTOS, N., WANG, D., YANG, H., HUYNH, D., VELAZQUEZ, L. & FIGUEROA, K. P. 2005. Spinocerebellar ataxia type 2: polyQ repeat variation in the CACNA1A calcium channel modifies age of onset. *Brain,* 128**,** 2297-303.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

PURCELL, S. M., WRAY, N. R., STONE, J. L., VISSCHER, P. M., O'DONOVAN, M. C., SULLIVAN, P. F. & SKLAR, P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature,* 460**,** 748-52.

QUARRELL, O. W., HANDLEY, O., O'DONOVAN, K., DUMOULIN, C., RAMOS-ARROYO, M., BIUNNO, I., BAUER, P., KLINE, M., LANDWEHRMEYER, G. B. & EUROPEAN HUNTINGTON'S DISEASE, N. 2012. Discrepancies in reporting the CAG repeat lengths for Huntington's disease. *Eur J Hum Genet,* 20**,** 20-6.

RAMPERSAD, R. R., TARRANT, T. K., VALLANAT, C. T., QUINTERO-MATTHEWS, T., WEEKS, M. F., ESSERMAN, D. A., CLARK, J., DI PADOVA, F., PATEL, D. D., FONG, A. M. & LIU, P. 2011. Enhanced Th17-cell responses render CCR2-deficient mice more susceptible for autoimmune arthritis. *PLoS One,* 6**,** e25833.

RANUM, L. P., CHUNG, M. Y., BANFI, S., BRYER, A., SCHUT, L. J., RAMESAR, R., DUVICK, L. A., MCCALL, A., SUBRAMONY, S. H., GOLDFARB, L. & ET AL. 1994. Molecular and clinical correlations in spinocerebellar ataxia type I: evidence for familial effects on the age at onset. *Am J Hum Genet,* 55**,** 244-52.

RAO, T., LONGERICH, S., ZHAO, W., AIHARA, H., SUNG, P. & XIONG, Y. 2018. Importance of homo-dimerization of Fanconi-associated nuclease 1 in DNA flap cleavage. *DNA Repair (Amst),* 64**,** 53-58.

RASCHLE, M., KNIPSCHEER, P., ENOIU, M., ANGELOV, T., SUN, J., GRIFFITH, J. D., ELLENBERGER, T. E., SCHARER, O. D. & WALTER, J. C. 2008. Mechanism of replication-coupled DNA interstrand crosslink repair. *Cell,* 134**,** 969-80.

RAWLINS, M. 2010. Huntington's disease out of the closet? *Lancet,* 376**,** 1372-3.

RAY CHAUDHURI, A., HASHIMOTO, Y., HERRADOR, R., NEELSEN, K. J., FACHINETTI, D., BERMEJO, R., COCITO, A., COSTANZO, V. & LOPES, M. 2012. Topoisomerase I poisoning results in PARP-mediated replication fork reversal. *Nat Struct Mol Biol,* 19**,** 417-23.

REACTOME. 2016. *Reactome Pathway Database* [Online]. Available: http://www.reactome.org/ [Accessed].

REDDY, K., SCHMIDT, M. H., GEIST, J. M., THAKKAR, N. P., PANIGRAHI, G. B., WANG, Y. H. & PEARSON, C. E. 2014. Processing of double-R-loops in (CAG).(CTG) and C9orf72 (GGGGCC).(GGCCCC) repeats causes instability. *Nucleic Acids Res,* 42**,** 10473-87.

REDDY, P. H. & SHIRENDEB, U. P. 2012. Mutant huntingtin, abnormal mitochondrial dynamics, defective axonal transport of mitochondria, and selective synaptic degeneration in Huntington's disease. *Biochim Biophys Acta,* 1822**,** 101-10.

REILMANN, R., SQUITIERI, F., PRILLER, J., SAFT, C., MARIOTTI, C., SUESSMUTH, S., NEMETH, A., TABRIZI, S., QUARRELL, O. & CRAUFURD, D. 2014. Safety and tolerability of selisistat for the treatment of Huntington's disease: results from a randomized, double-blind, placebo-controlled phase II trial (S47. 004). *Neurology,* 82**,** S47. 004-S47. 004.

REINER, A., ALBIN, R. L., ANDERSON, K. D., D'AMATO, C. J., PENNEY, J. B. & YOUNG, A. B. 1988. Differential loss of striatal projection neurons in Huntington disease. *Proc Natl Acad Sci U S A,* 85**,** 5733-7.

REN, Y., LAI, Y., LAVERDE, E. E., LEI, R., REIN, H. L. & LIU, Y. 2017. Modulation of trinucleotide repeat instability by DNA polymerase beta polymorphic variant R137Q. *PLoS One,* 12**,** e0177299.

RENNA, M., JIMENEZ-SANCHEZ, M., SARKAR, S. & RUBINSZTEIN, D. C. 2010. Chemical Inducers of Autophagy That Enhance the Clearance of Mutant Proteins in Neurodegenerative Diseases. *Journal of Biological Chemistry,* 285**,** 11061-11067.

REYNIERS, E., MARTIN, J. J., CRAS, P., VAN MARCK, E., HANDIG, I., JORENS, H. Z., OOSTRA, B. A., KOOY, R. F. & WILLEMS, P. J. 1999. Postmortem examination of two fragile X brothers with an FMR1 full mutation. *Am J Med Genet,* 84**,** 245-9.

RHODES, L. E., FREEMAN, B. K., AUH, S., KOKKINIS, A. D., LA PEAN, A., CHEN, C., LEHKY, T. J., SHRADER, J. A., LEVY, E. W., HARRIS-LOVE, M., DI PROSPERO, N. A. & FISCHBECK, K. H. 2009. Clinical features of spinal and bulbar muscular atrophy. *Brain,* 132**,** 3242-51.

RIESS, O., RUB, U., PASTORE, A., BAUER, P. & SCHOLS, L. 2008. SCA3: neurological features, pathogenesis and animal models. *Cerebellum,* 7**,** 125-37.

RISINGER, J. I., UMAR, A., BOYD, J., BERCHUCK, A., KUNKEL, T. A. & BARRETT, J. C. 1996. Mutation of MSH3 in endometrial cancer and evidence for its functional role in heteroduplex repair. *Nat Genet,* 14**,** 102-5.

RODRIGUEZ, G. P., ROMANOVA, N. V., BAO, G., ROUF, N. C., KOW, Y. W. & CROUSE, G. F. 2012. Mismatch repair-dependent mutagenesis in nondividing cells. *Proc Natl Acad Sci U S A,* 109**,** 6153-8.

ROLFS, A., KOEPPEN, A. H., BAUER, I., BAUER, P., BUHLMANN, S., TOPKA, H., SCHOLS, L. & RIESS, O. 2003. Clinical features and neuropathology of autosomal dominant spinocerebellar ataxia (SCA17). *Ann Neurol,* 54**,** 367-75.

ROLSETH, V., KROKEIDE, S. Z., KUNKE, D., NEURAUTER, C. G., SUGANTHAN, R., SEJERSTED, Y., HILDRESTRAND, G. A., BJORAS, M. & LUNA, L. 2013. Loss of Neil3, the major DNA glycosylase activity for removal of hydantoins in single stranded DNA, reduces cellular proliferation and sensitizes cells to genotoxic stress. *Biochim Biophys Acta,* 1833**,** 1157-64.

ROSS, C. A., AYLWARD, E. H., WILD, E. J., LANGBEHN, D. R., LONG, J. D., WARNER, J. H., SCAHILL, R. I., LEAVITT, B. R., STOUT, J. C., PAULSEN, J. S., REILMANN, R., UNSCHULD, P. G., WEXLER, A., MARGOLIS, R. L. & TABRIZI, S. J. 2014. Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nat Rev Neurol,* 10**,** 204-16.

ROSS, C. A. & TABRIZI, S. J. 2011. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet Neurol,* 10**,** 83-98.

ROSS, C. A. & TRUANT, R. 2017. DNA repair: A unifying mechanism in neurodegeneration. *Nature,* 541**,** 34-35.

ROSSER, A. E. & BACHOUD-LEVI, A. C. 2012. Clinical trials of neural transplantation in Huntington's disease. *Prog Brain Res,* 200**,** 345-71.

ROTHFUSS, A. & GROMPE, M. 2004. Repair kinetics of genomic interstrand DNA cross-links: evidence for DNA double-strand break-dependent activation of the Fanconi anemia/BRCA pathway. *Mol Cell Biol,* 24**,** 123-34.

RUBINSZTEIN, D. C., LEGGO, J., CHIANO, M., DODGE, A., NORBURY, G., ROSSER, E. & CRAUFURD, D. 1997. Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. *Proc Natl Acad Sci U S A,* 94**,** 3872-6.

RUNNE, H., KUHN, A., WILD, E. J., PRATYAKSHA, W., KRISTIANSEN, M., ISAACS, J. D., REGULIER, E., DELORENZI, M., TABRIZI, S. J. & LUTHI-CARTER, R. 2007. Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood. *Proc Natl Acad Sci U S A,* 104**,** 14424-9.

SADRI-VAKILI, G., BOUZOU, B., BENN, C. L., KIM, M. O., CHAWLA, P., OVERLAND, R. P., GLAJCH, K. E., XIA, E., QIU, Z., HERSCH, S. M., CLARK, T. W., YOHRLING, G. J. & CHA, J. H. 2007. Histones

associated with downregulated genes are hypo-acetylated in Huntington's disease models. *Hum Mol Genet,* 16**,** 1293-306.

SALEH, N., MOUTEREAU, S., DURR, A., KRYSTKOWIAK, P., AZULAY, J. P., TRANCHANT, C., BROUSSOLLE, E., MORIN, F., BACHOUD-LEVI, A. C. & MAISON, P. 2009. Neuroendocrine disturbances in Huntington's disease. *PLoS One,* 4**,** e4962.

SANDERSON, B. J. & SHIELD, A. J. 1996. Mutagenic damage to mammalian cells by therapeutic alkylating agents. *Mutat Res,* 355**,** 41-57.

SAPP, E., KEGEL, K. B., ARONIN, N., HASHIKAWA, T., UCHIYAMA, Y., TOHYAMA, K., BHIDE, P. G., VONSATTEL, J. P. & DIFIGLIA, M. 2001. Early and progressive accumulation of reactive microglia in the Huntington disease brain. *J Neuropathol Exp Neurol,* 60**,** 161-72.

SAVAS, J. N., MAKUSKY, A., OTTOSEN, S., BAILLAT, D., THEN, F., KRAINC, D., SHIEKHATTAR, R., MARKEY, S. P. & TANESE, N. 2008. Huntington's disease protein contributes to RNA-mediated gene silencing through association with Argonaute and P bodies. *Proc Natl Acad Sci U S A,* 105**,** 10820-5.

SAVOURET, C., BRISSON, E., ESSERS, J., KANAAR, R., PASTINK, A., TE RIELE, H., JUNIEN, C. & GOURDON, G. 2003. CTG repeat instability and size variation timing in DNA repair-deficient mice. *EMBO J,* 22**,** 2264-73.

SAVOURET, C., GARCIA-CORDIER, C., MEGRET, J., TE RIELE, H., JUNIEN, C. & GOURDON, G. 2004. MSH2-dependent germinal CTG repeat expansions are produced continuously in spermatogonia from DM1 transgenic mice. *Mol Cell Biol,* 24**,** 629-37.

SCHLACHER, K., WU, H. & JASIN, M. 2012. A distinct replication fork protection pathway connects Fanconi anemia tumor suppressors to RAD51-BRCA1/2. *Cancer Cell,* 22**,** 106-16.

SCHMIDT, M. H. & PEARSON, C. E. 2016. Disease-associated repeat instability and mismatch repair. *DNA Repair (Amst),* 38**,** 117-26.

SCHMUTTE, C., SADOFF, M. M., SHIM, K. S., ACHARYA, S. & FISHEL, R. 2001. The interaction of DNA mismatch repair proteins with human exonuclease I. *J Biol Chem,* 276**,** 33011-8.

SCHNEIDER, S. A., VAN DE WARRENBURG, B. P., HUGHES, T. D., DAVIS, M., SWEENEY, M., WOOD, N., QUINN, N. P. & BHATIA, K. P. 2006. Phenotypic homogeneity of the Huntington disease-like presentation in a SCA17 family. *Neurology,* 67**,** 1701-3.

SCHOFIELD, M. J. & HSIEH, P. 2003. DNA mismatch repair: molecular mechanisms and biological function. *Annu Rev Microbiol,* 57**,** 579-608.

SCIENTIFIC, T. 2018. *ChIP Analysis - UK* [Online]. Available: https://www.thermofisher.com/uk/en/home/life-science/epigenetics-noncoding-rna-research/chromatin-remodeling/chromatin-immunoprecipitation-chip/chip-analysis.html [Accessed 24/08/2018].

SEGUI, N., MINA, L. B., LAZARO, C., SANZ-PAMPLONA, R., PONS, T., NAVARRO, M., BELLIDO, F., LOPEZ-DORIGA, A., VALDES-MAS, R., PINEDA, M., GUINO, E., VIDAL, A., SOTO, J. L., CALDES, T., DURAN, M., URIOSTE, M., RUEDA, D., BRUNET, J., BALBIN, M., BLAY, P., IGLESIAS, S., GARRE, P., LASTRA, E., SANCHEZ-HERAS, A. B., VALENCIA, A., MORENO, V., PUJANA, M. A., VILLANUEVA, A., BLANCO, I., CAPELLA, G., SURRALLES, J., PUENTE, X. S. & VALLE, L. 2015a. Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. *Gastroenterology,* 149**,** 563-6.

SEGUI, N., MINA, L. B., LAZARO, C., SANZ-PAMPLONA, R., PONS, T., NAVARRO, M., BELLIDO, F., LOPEZ-DORIGA, A., VALDES-MAS, R., PINEDA, M., GUINO, E., VIDAL, A., SOTO, J. L., CALDES, T., DURAN, M., URIOSTE, M., RUEDA, D., BRUNET, J., BALBIN, M., BLAY, P., IGLESIAS, S., GARRE, P., LASTRA, E., SANCHEZ-HERAS, A. B., VALENCIA, A., MORENO, V., PUJANA, M. A., VILLANUEVA, A., BLANCO, I., CAPELLA, G., SURRALLES, J., PUENTE, X. S. & VALLE, L. 2015b.

Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. *Gastroenterology*.

SEREDENINA, T. & LUTHI-CARTER, R. 2012. What have we learned from gene expression profiles in Huntington's disease? *Neurobiol Dis,* 45**,** 83-98.

SERIOLA, A., SPITS, C., SIMARD, J. P., HILVEN, P., HAENTJENS, P., PEARSON, C. E. & SERMON, K. 2011a. Huntington's and myotonic dystrophy hESCs: down-regulated trinucleotide repeat instability and mismatch repair machinery expression upon differentiation. *Human Molecular Genetics,* 20**,** 176-85.

SERIOLA, A., SPITS, C., SIMARD, J. P., HILVEN, P., HAENTJENS, P., PEARSON, C. E. & SERMON, K. 2011b. Huntington's and myotonic dystrophy hESCs: down-regulated trinucleotide repeat instability and mismatch repair machinery expression upon differentiation. *Hum Mol Genet,* 20**,** 176-85.

SEZNEC, H., LIA-BALDINI, A. S., DUROS, C., FOUQUET, C., LACROIX, C., HOFMANN-RADVANYI, H., JUNIEN, C. & GOURDON, G. 2000. Transgenic mice carrying large human genomic sequences with expanded CTG repeat mimic closely the DM CTG repeat intergenerational and somatic instability. *Hum Mol Genet,* 9**,** 1185-94.

SHAH, K. A. & MIRKIN, S. M. 2015. The hidden side of unstable DNA repeats: Mutagenesis at a distance. *DNA Repair (Amst),* 32**,** 106-12.

SHELBOURNE, P. F., KELLER-MCGANDY, C., BI, W. L., YOON, S. R., DUBEAU, L., VEITCH, N. J., VONSATTEL, J. P., WEXLER, N. S., ARNHEIM, N. & AUGOOD, S. J. 2007a. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. *Hum Mol Genet,* 16**,** 1133-42.

SHELBOURNE, P. F., KELLER-MCGANDY, C., BI, W. L., YOON, S. R., DUBEAU, L., VEITCH, N. J., VONSATTEL, J. P., WEXLER, N. S., GROUP, U. S.-V. C. R., ARNHEIM, N. & AUGOOD, S. J. 2007b. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. *Hum Mol Genet,* 16**,** 1133-42.

SHEREDA, R. D., MACHIDA, Y. & MACHIDA, Y. J. 2010. Human KIAA1018/FAN1 localizes to stalled replication forks via its ubiquitin-binding domain. *Cell Cycle,* 9**,** 3977-83.

SHI, Y., KIRWAN, P. & LIVESEY, F. J. 2012. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nat Protoc,* 7**,** 1836-46.

SHILOH, Y. & ZIV, Y. 2013. The ATM protein kinase: regulating the cellular response to genotoxic stress, and more. *Nat Rev Mol Cell Biol,* 14**,** 197-210.

SHIMOHATA, T., NAKAJIMA, T., YAMADA, M., UCHIDA, C., ONODERA, O., NARUSE, S., KIMURA, T., KOIDE, R., NOZAKI, K., SANO, Y., ISHIGURO, H., SAKOE, K., OOSHIMA, T., SATO, A., IKEUCHI, T., OYAKE, M., SATO, T., AOYAGI, Y., HOZUMI, I., NAGATSU, T., TAKIYAMA, Y., NISHIZAWA, M., GOTO, J., KANAZAWA, I., DAVIDSON, I., TANESE, N., TAKAHASHI, H. & TSUJI, S. 2000. Expanded polyglutamine stretches interact with TAFII130, interfering with CREB-dependent transcription. *Nat Genet,* 26**,** 29-36.

SILVEIRA, I., MIRANDA, C., GUIMARAES, L., MOREIRA, M. C., ALONSO, I., MENDONCA, P., FERRO, A., PINTO-BASTO, J., COELHO, J., FERREIRINHA, F., POIRIER, J., PARREIRA, E., VALE, J., JANUARIO, C., BARBOT, C., TUNA, A., BARROS, J., KOIDE, R., TSUJI, S., HOLMES, S. E., MARGOLIS, R. L., JARDIM, L., PANDOLFO, M., COUTINHO, P. & SEQUEIROS, J. 2002. Trinucleotide repeats in 202 families with ataxia: a small expanded (CAG)n allele at the SCA17 locus. *Arch Neurol,* 59**,** 623-9.

SIMMONS, D. A., BELICHENKO, N. P., YANG, T., CONDON, C., MONBUREAU, M., SHAMLOO, M., JING, D., MASSA, S. M. & LONGO, F. M. 2013. A Small Molecule TrkB Ligand Reduces Motor Impairment and Neuropathology in R6/2 and BACHD Mouse Models of Huntington's Disease. *The Journal of Neuroscience,* 33**,** 18712-18727.

SIMMONS, D. A., CASALE, M., ALCON, B., PHAM, N., NARAYAN, N. & LYNCH, G. 2007. Ferritin accumulation in dystrophic microglia is an early event in the development of Huntington's disease. *Glia,* 55**,** 1074-84.

SINGHRAO, S. K., NEAL, J. W., MORGAN, B. P. & GASQUE, P. 1999. Increased complement biosynthesis by microglia and complement activation on neurons in Huntington's disease. *Exp Neurol,* 159**,** 362-76.

SINNREICH, M., SORENSON, E. J. & KLEIN, C. J. 2004. Neurologic course, endocrine dysfunction and triplet repeat size in spinal bulbar muscular atrophy. *Can J Neurol Sci,* 31**,** 378-82.

SLEAN, M. M., PANIGRAHI, G. B., CASTEL, A. L., PEARSON, A. B., TOMKINSON, A. E. & PEARSON, C. E. 2016. Absence of MutSbeta leads to the formation of slipped-DNA for CTG/CAG contractions at primate replication forks. *DNA Repair (Amst),* 42**,** 107-18.

SLEAN, M. M., PANIGRAHI, G. B., RANUM, L. P. & PEARSON, C. E. 2008. Mutagenic roles of DNA "repair" proteins in antibody diversity and disease-associated trinucleotide repeat instability. *DNA Repair,* 7**,** 1135-1154.

SMITH, A. L., ALIREZAIE, N., CONNOR, A., CHAN-SENG-YUE, M., GRANT, R., SELANDER, I., BASCUNANA, C., BORGIDA, A., HALL, A., WHELAN, T., HOLTER, S., MCPHERSON, T., CLEARY, S., PETERSEN, G. M., OMEROGLU, A., SALOUSTROS, E., MCPHERSON, J., STEIN, L. D., FOULKES, W. D., MAJEWSKI, J., GALLINGER, S. & ZOGOPOULOS, G. 2016. Candidate DNA repair susceptibility genes identified by exome sequencing in high-risk pancreatic cancer. *Cancer Lett,* 370**,** 302-12.

SMITH, M. R., SYED, A., LUKACSOVICH, T., PURCELL, J., BARBARO, B. A., WORTHGE, S. A., WEI, S. R., POLLIO, G., MAGNONI, L., SCALI, C., MASSAI, L., FRANCESCHINI, D., CAMARRI, M., GIANFRIDDO, M., DIODATO, E., THOMAS, R., GOKCE, O., TABRIZI, S. J., CARICASOLE, A., LANDWEHRMEYER, B., MENALLED, L., MURPHY, C., RAMBOZ, S., LUTHI-CARTER, R., WESTERBERG, G. & MARSH, J. L. 2014. A potent and selective Sirtuin 1 inhibitor alleviates pathology in multiple animal and cell models of Huntington's disease. *Hum Mol Genet,* 23**,** 2995-3007.

SMOGORZEWSKA, A., DESETTY, R., SAITO, T. T., SCHLABACH, M., LACH, F. P., SOWA, M. E., CLARK, A. B., KUNKEL, T. A., HARPER, J. W., COLAIACOVO, M. P. & ELLEDGE, S. J. 2010a. A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. *Mol Cell,* 39**,** 36-47.

SMOGORZEWSKA, A., DESETTY, R., SAITO, T. T., SCHLABACH, M., LACH, F. P., SOWA, M. E., CLARK, A. B., KUNKEL, T. A., HARPER, J. W., COLAIACOVO, M. P. & ELLEDGE, S. J. 2010b. A genetic screen identifies FAN1, a Fanconi anemia-associated nuclease necessary for DNA interstrand crosslink repair. *Mol Cell.* United States: 2010 Elsevier Inc.

SNELL, R. G., MACMILLAN, J. C., CHEADLE, J. P., FENTON, I., LAZAROU, L. P., DAVIES, P., MACDONALD, M. E., GUSELLA, J. F., HARPER, P. S. & SHAW, D. J. 1993. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nat Genet,* 4**,** 393-7.

SOBCZAK, K. & KRZYZOSIAK, W. J. 2004. Imperfect CAG repeats form diverse structures in SCA1 transcripts. *J Biol Chem,* 279**,** 41563-72.

SOKOLOVSKY, N., COOK, A., HUNT, H., GIUNTI, P. & CIPOLOTTI, L. 2010. A preliminary characterisation of cognition and social cognition in spinocerebellar ataxia types 2, 1, and 7. *Behav Neurol,* 23**,** 17-29.

SONTAG, E. M., JOACHIMIAK, L. A., TAN, Z., TOMLINSON, A., HOUSMAN, D. E., GLABE, C. G., POTKIN, S. G., FRYDMAN, J. & THOMPSON, L. M. 2013. Exogenous delivery of chaperonin subunit fragment ApiCCT1 modulates mutant Huntingtin cellular phenotypes. *Proceedings of the National Academy of Sciences,* 110**,** 3077-3082.

SPADA, A. L. 2014. Spinal and Bulbar Muscular Atrophy.

SPAIN, S. L. & BARRETT, J. C. 2015. Strategies for fine-mapping complex traits. *Hum Mol Genet,* 24**,** R111-9.

SPENCER, J. P., JENNER, A., ARUOMA, O. I., CROSS, C. E., WU, R. & HALLIWELL, B. 1996. Oxidative DNA damage in human respiratory tract epithelial cells. Time course in relation to DNA strand breakage. *Biochem Biophys Res Commun,* 224**,** 17-22.

SPENCER, J. P., JENNER, A., CHIMEL, K., ARUOMA, O. I., CROSS, C. E., WU, R. & HALLIWELL, B. 1995. DNA strand breakage and base modification induced by hydrogen peroxide treatment of human respiratory tract epithelial cells. *FEBS Lett,* 374**,** 233-6.

SRINIVASAN, K., FRIEDMAN, B. A., LARSON, J. L., LAUFFER, B. E., GOLDSTEIN, L. D., APPLING, L. L., BORNEO, J., POON, C., HO, T., CAI, F., STEINER, P., VAN DER BRUG, M. P., MODRUSAN, Z., KAMINKER, J. S. & HANSEN, D. V. 2016. Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. *Nat Commun,* 7**,** 11295.

STEFFAN, J. S., BODAI, L., PALLOS, J., POELMAN, M., MCCAMPBELL, A., APOSTOL, B. L., KAZANTSEV, A., SCHMIDT, E., ZHU, Y. Z., GREENWALD, M., KUROKAWA, R., HOUSMAN, D. E., JACKSON, G. R., MARSH, J. L. & THOMPSON, L. M. 2001. Histone deacetylase inhibitors arrest polyglutamine-dependent neurodegeneration in Drosophila. *Nature,* 413**,** 739-43.

STEVENS, J. R., LAHUE, E. E., LI, G. M. & LAHUE, R. S. 2013. Trinucleotide repeat expansions catalyzed by human cell-free extracts. *Cell Res,* 23**,** 565-72.

STOREY, E., FORREST, S. M., SHAW, J. H., MITCHELL, P. & GARDNER, R. J. 1999. Spinocerebellar ataxia type 2: clinical features of a pedigree displaying prominent frontal-executive dysfunction. *Arch Neurol,* 56**,** 43-50.

STOREY, J. D. & TIBSHIRANI, R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A,* 100**,** 9440-5.

STRACCIA, M., GARCIA-DIAZ BARRIGA, G., SANDERS, P., BOMBAU, G., CARRERE, J., MAIRAL, P. B., VINH, N. N., YUNG, S., KELLY, C. M., SVENDSEN, C. N., KEMP, P. J., ARJOMAND, J., SCHOENFELD, R. C., ALBERCH, J., ALLEN, N. D., ROSSER, A. E. & CANALS, J. M. 2015. Quantitative high-throughput gene expression profiling of human striatal development to screen stem cell-derived medium spiny neurons. *Mol Ther Methods Clin Dev,* 2**,** 15030.

STRAND, A. D., ARAGAKI, A. K., SHAW, D., BIRD, T., HOLTON, J., TURNER, C., TAPSCOTT, S. J., TABRIZI, S. J., SCHAPIRA, A. H., KOOPERBERG, C. & OLSON, J. M. 2005. Gene expression in Huntington's disease skeletal muscle: a potential biomarker. *Hum Mol Genet,* 14**,** 1863-76.

SU, X. A. & FREUDENREICH, C. H. 2017. Cytosine deamination and base excision repair cause R-loop-induced CAG repeat fragility and instability in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A,* 114**,** E8392-E8401.

SUBERBIELLE, E., SANCHEZ, P. E., KRAVITZ, A. V., WANG, X., HO, K., EILERTSON, K., DEVIDZE, N., KREITZER, A. C. & MUCKE, L. 2013. Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-beta. *Nat Neurosci,* 16**,** 613-21.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

SUBRAMONY, S. 2012. The Ataxias. *In:* WOOD, N. W. (ed.) *Neurogenetics. A guide for Clinicians.* Cambridge University Press.

SUNKIN, S. M., NG, L., LAU, C., DOLBEARE, T., GILBERT, T. L., THOMPSON, C. L., HAWRYLYCZ, M. & DANG, C. 2013. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res,* 41**,** D996-D1008.

SVEINBJORNSSON, G., ALBRECHTSEN, A., ZINK, F., GUDJONSSON, S. A., ODDSON, A., MASSON, G., HOLM, H., KONG, A., THORSTEINSDOTTIR, U., SULEM, P., GUDBJARTSSON, D. F. & STEFANSSON, K. 2016. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet,* 48**,** 314-7.

SWAMI, M., HENDRICKS, A. E., GILLIS, T., MASSOOD, T., MYSORE, J., MYERS, R. H. & WHEELER, V. C. 2009. Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum Mol Genet,* 18**,** 3039-47.

SWANN, P. F., WATERS, T. R., MOULTON, D. C., XU, Y. Z., ZHENG, Q., EDWARDS, M. & MACE, R. 1996. Role of postreplicative DNA mismatch repair in the cytotoxic action of thioguanine. *Science,* 273**,** 1109-11.

TABRIZI, S. J., LANGBEHN, D. R., LEAVITT, B. R., ROOS, R. A., DURR, A., CRAUFURD, D., KENNARD, C., HICKS, S. L., FOX, N. C., SCAHILL, R. I., BOROWSKY, B., TOBIN, A. J., ROSAS, H. D., JOHNSON, H., REILMANN, R., LANDWEHRMEYER, B. & STOUT, J. C. 2009a. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet Neurol,* 8**,** 791-801.

TABRIZI, S. J., LANGBEHN, D. R., LEAVITT, B. R., ROOS, R. A., DURR, A., CRAUFURD, D., KENNARD, C., HICKS, S. L., FOX, N. C., SCAHILL, R. I., BOROWSKY, B., TOBIN, A. J., ROSAS, H. D., JOHNSON, H., REILMANN, R., LANDWEHRMEYER, B., STOUT, J. C. & INVESTIGATORS, T.-H. 2009b. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet Neurol,* 8**,** 791-801.

TABRIZI, S. J., REILMANN, R., ROOS, R. A., DURR, A., LEAVITT, B., OWEN, G., JONES, R., JOHNSON, H., CRAUFURD, D., HICKS, S. L., KENNARD, C., LANDWEHRMEYER, B., STOUT, J. C., BOROWSKY, B., SCAHILL, R. I., FROST, C. & LANGBEHN, D. R. 2012. Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. *Lancet Neurol,* 11**,** 42-53.

TABRIZI, S. J., SCAHILL, R. I., DURR, A., ROOS, R. A., LEAVITT, B. R., JONES, R., LANDWEHRMEYER, G. B., FOX, N. C., JOHNSON, H., HICKS, S. L., KENNARD, C., CRAUFURD, D., FROST, C., LANGBEHN, D. R., REILMANN, R. & STOUT, J. C. 2011a. Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol,* 10**,** 31-42.

TABRIZI, S. J., SCAHILL, R. I., DURR, A., ROOS, R. A., LEAVITT, B. R., JONES, R., LANDWEHRMEYER, G. B., FOX, N. C., JOHNSON, H., HICKS, S. L., KENNARD, C., CRAUFURD, D., FROST, C., LANGBEHN, D. R., REILMANN, R., STOUT, J. C. & INVESTIGATORS, T.-H. 2011b. Biological and clinical changes in premanifest and early stage Huntington's disease in the TRACK-HD study: the 12-month longitudinal analysis. *Lancet Neurol,* 10**,** 31-42.

TABRIZI, S. J., SCAHILL, R. I., OWEN, G., DURR, A., LEAVITT, B. R., ROOS, R. A., BOROWSKY, B., LANDWEHRMEYER, B., FROST, C., JOHNSON, H., CRAUFURD, D., REILMANN, R., STOUT, J. C. & LANGBEHN, D. R. 2013. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. *Lancet Neurol,* 12**,** 637-49.

TAHERZADEH-FARD, E., SAFT, C., ANDRICH, J., WIECZOREK, S. & ARNING, L. 2009. PGC-1alpha as modifier of onset age in Huntington disease. *Mol Neurodegener,* 4**,** 10.

TAI, Y. F., PAVESE, N., GERHARD, A., TABRIZI, S. J., BARKER, R. A., BROOKS, D. J. & PICCINI, P. 2007a. Microglial activation in presymptomatic Huntington's disease gene carriers. *Brain,* 130**,** 1759-1766.

TAI, Y. F., PAVESE, N., GERHARD, A., TABRIZI, S. J., BARKER, R. A., BROOKS, D. J. & PICCINI, P. 2007b. Microglial activation in presymptomatic Huntington's disease gene carriers. *Brain,* 130**,** 1759-66.

TAKAHASHI, D., SATO, K., HIRAYAMA, E., TAKATA, M. & KURUMIZAKA, H. 2015. Human FAN1 promotes strand incision in 5'-flapped DNA complexed with RPA. *J Biochem,* 158**,** 263-70.

TAKANO, H., ONODERA, O., TAKAHASHI, H., IGARASHI, S., YAMADA, M., OYAKE, M., IKEUCHI, T., KOIDE, R., TANAKA, H., IWABUCHI, K. & TSUJI, S. 1996. Somatic mosaicism of expanded CAG repeats in brains of patients with dentatorubral-pallidoluysian atrophy: cellular population-dependent dynamics of mitotic instability. *Am J Hum Genet,* 58**,** 1212-22.

TANAKA, F., REEVES, M. F., ITO, Y., MATSUMOTO, M., LI, M., MIWA, S., INUKAI, A., YAMAMOTO, M., DOYU, M., YOSHIDA, M., HASHIZUME, Y., TERAO, S., MITSUMA, T. & SOBUE, G. 1999. Tissue-specific somatic mosaicism in spinal and bulbar muscular atrophy is dependent on CAG-repeat length and androgen receptor--gene expression level. *Am J Hum Genet,* 65**,** 966-73.

TAYLOR, A. K., TASSONE, F., DYER, P. N., HERSCH, S. M., HARRIS, J. B., GREENOUGH, W. T. & HAGERMAN, R. J. 1999. Tissue heterogeneity of the FMR1 mutation in a high-functioning male with fragile X syndrome. *Am J Med Genet,* 84**,** 233-9.

TAYLOR, D. M., MOSER, R., RÉGULIER, E., BREUILLAUD, L., DIXON, M., BEESEN, A. A., ELLISTON, L., SILVA SANTOS, M. D. F., KIM, J., JONES, L., GOLDSTEIN, D. R., FERRANTE, R. J. & LUTHI-CARTER, R. 2013. MAP Kinase Phosphatase 1 (MKP-1/DUSP1) Is Neuroprotective in Huntington's Disease via Additive Effects of JNK and p38 Inhibition. *The Journal of Neuroscience,* 33**,** 2313-2325.

TEAM, R. C. 2013. *R: A language and environment for statistical computing* [Online]. R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.R-project.org/ [Accessed 27/08/2018].

TELENIUS, H., ALMQVIST, E., KREMER, B., SPENCE, N., SQUITIERI, F., NICHOL, K., GRANDELL, U., STARR, E., BENJAMIN, C., CASTALDO, I. & ET AL. 1995. Somatic mosaicism in sperm is associated with intergenerational (CAG)n changes in Huntington disease. *Hum Mol Genet,* 4**,** 189-95.

TELENIUS, H., KREMER, B., GOLDBERG, Y. P., THEILMANN, J., ANDREW, S. E., ZEISLER, J., ADAM, S., GREENBERG, C., IVES, E. J., CLARKE, L. A. & ET AL. 1994. Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nat Genet,* 6**,** 409-14.

TEO, C. R., WANG, W., YANG LAW, H., LEE, C. G. & CHONG, S. S. 2008. Single-step scalable-throughput molecular screening for Huntington disease. *Clin Chem,* 54**,** 964-72.

TEZENAS DU MONTCEL, S., DURR, A., BAUER, P., FIGUEROA, K. P., ICHIKAWA, Y., BRUSSINO, A., FORLANI, S., RAKOWICZ, M., SCHOLS, L., MARIOTTI, C., VAN DE WARRENBURG, B. P., ORSI, L., GIUNTI, P., FILLA, A., SZYMANSKI, S., KLOCKGETHER, T., BERCIANO, J., PANDOLFO, M., BOESCH, S., MELEGH, B., TIMMANN, D., MANDICH, P., CAMUZAT, A., GOTO, J., ASHIZAWA, T., CAZENEUVE, C., TSUJI, S., PULST, S. M., BRUSCO, A., RIESS, O., BRICE, A. & STEVANIN, G. 2014. Modulation of the age at onset in spinocerebellar ataxia by CAG tracts in various genes. *Brain,* 137**,** 2444-55.

THOMAS, E. A., COPPOLA, G., DESPLATS, P. A., TANG, B., SORAGNI, E., BURNETT, R., GAO, F., FITZGERALD, K. M., BOROK, J. F., HERMAN, D., GESCHWIND, D. H. & GOTTESFELD, J. M. 2008. The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in Huntington's disease transgenic mice. *Proc Natl Acad Sci U S A,* 105**,** 15564-9.

THONGTHIP, S., BELLANI, M., GREGG, S. Q., SRIDHAR, S., CONTI, B. A., CHEN, Y., SEIDMAN, M. M. & SMOGORZEWSKA, A. 2016. Fan1 deficiency results in DNA interstrand cross-link repair defects, enhanced tissue karyomegaly, and organ dysfunction. *Genes Dev,* 30**,** 645-59.

THORNTON, C. A. 2014. Myotonic Dystrophy. *Neurol Clin,* 32**,** 705-19.

THORNTON, C. A., JOHNSON, K. & MOXLEY, R. T., 3RD 1994. Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. *Ann Neurol,* 35**,** 104-7.

TODD, D., GOWERS, I., DOWLER, S. J., WALL, M. D., MCALLISTER, G., FISCHER, D. F., DIJKSTRA, S., FRATANTONI, S. A., VAN DE BOSPOORT, R., VEENMAN-KOEPKE, J., FLYNN, G., ARJOMAND, J., DOMINGUEZ, C., MUNOZ-SANJUAN, I., WITYAK, J. & BARD, J. A. 2014. A monoclonal antibody TrkB receptor agonist as a potential therapeutic for Huntington's disease. *PLoS One,* 9**,** e87923.

TOME, S., MANLEY, K., SIMARD, J. P., CLARK, G. W., SLEAN, M. M., SWAMI, M., SHELBOURNE, P. F., TILLIER, E. R., MONCKTON, D. G., MESSER, A. & PEARSON, C. E. 2013a. MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. *PLoS Genet,* 9**,** e1003280.

TOME, S., PANIGRAHI, G. B., LOPEZ CASTEL, A., FOIRY, L., MELTON, D. W., GOURDON, G. & PEARSON, C. E. 2011. Maternal germline-specific effect of DNA ligase I on CTG/CAG instability. *Human Molecular Genetics,* 20**,** 2131-2143.

TOME, S., SIMARD, J. P., SLEAN, M. M., HOLT, I., MORRIS, G. E., WOJCIECHOWICZ, K., TE RIELE, H. & PEARSON, C. E. 2013b. Tissue-specific mismatch repair protein expression: MSH3 is higher than MSH6 in multiple mouse tissues. *DNA Repair (Amst),* 12**,** 46-52.

TOMITA, H., VAWTER, M. P., WALSH, D. M., EVANS, S. J., CHOUDARY, P. V., LI, J., OVERMAN, K. M., ATZ, M. E., MYERS, R. M., JONES, E. G., WATSON, S. J., AKIL, H. & BUNNEY, W. E., JR. 2004. Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain. *Biol Psychiatry,* 55**,** 346-52.

TRAGER, U., ANDRE, R., LAHIRI, N., MAGNUSSON-LIND, A., WEISS, A., GRUENINGER, S., MCKINNON, C., SIRINATHSINGHJI, E., KAHLON, S., PFISTER, E. L., MOSER, R., HUMMERICH, H., ANTONIOU, M., BATES, G. P., LUTHI-CARTER, R., LOWDELL, M. W., BJORKQVIST, M., OSTROFF, G. R., ARONIN, N. & TABRIZI, S. J. 2014. HTT-lowering reverses Huntington's disease immune dysfunction caused by NFkappaB pathway dysregulation. *Brain,* 137**,** 819-33.

TRÄGER, U., ANDRE, R., MAGNUSSON-LIND, A., MILLER, J. R. C., CONNOLLY, C., WEISS, A., GRUENINGER, S., SILAJDŽIĆ, E., SMITH, D. L., LEAVITT, B. R., BATES, G. P., BJÖRKQVIST, M. & TABRIZI, S. J. 2015. Characterisation of immune cell function in fragment and full-length Huntington's disease mouse models. *Neurobiology of Disease,* 73**,** 388-398.

TRIALS, C. 2015. *Study Evaluating The Safety, Tolerability And Brain Function Of 2 Doses Of PF-0254920 In Subjects With Early Huntington's Disease - Full Text View - ClinicalTrials.gov* [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT01806896 [Accessed].

TRIALS, C. 2016. *Safety, Tolerability, Pharmacokinetics, and Pharmacodynamics of IONIS-HTTRx in Patients With Early Manifest Huntington's Disease - Full Text View - ClinicalTrials.gov* [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT02519036?term=huntington%27s+disease&rank=27 [Accessed].

TROTTIER, Y., DEVYS, D., IMBERT, G., SAUDOU, F., AN, I., LUTZ, Y., WEBER, C., AGID, Y., HIRSCH, E. C. & MANDEL, J.-L. 1995. Cellular localization of the Huntington's disease protein and discrimination of the normal and mutated form. *Nature Genetics,* 10**,** 104-110.

TURNER, C., COOPER, J. M. & SCHAPIRA, A. H. V. 2007. Clinical correlates of mitochondrial function in Huntington's disease muscle. *Movement Disorders,* 22**,** 1715-1721.

UDD, B., JUVONEN, V., HAKAMIES, L., NIEMINEN, A., WALLGREN-PETTERSSON, C., CEDERQUIST, K. & SAVONTAUS, M. L. 1998. High prevalence of Kennedy's disease in Western Finland -- is the syndrome underdiagnosed? *Acta Neurol Scand,* 98**,** 128-33.

UENO, S., KONDOH, K., KOTANI, Y., KOMURE, O., KUNO, S., KAWAI, J., HAZAMA, F. & SANO, A. 1995. Somatic mosaicism of CAG repeat in dentatorubral-pallidoluysian atrophy (DRPLA). *Hum Mol Genet,* 4**,** 663-6.

USDIN, K., HOUSE, N. C. & FREUDENREICH, C. H. 2015. Repeat instability during DNA repair: Insights from model systems. *Crit Rev Biochem Mol Biol,* 50**,** 142-67.

VAN DE WARRENBURG, B. P., HENDRIKS, H., DURR, A., VAN ZUIJLEN, M. C., STEVANIN, G., CAMUZAT, A., SINKE, R. J., BRICE, A. & KREMER, B. P. 2005. Age at onset variance analysis in spinocerebellar ataxias: a study in a Dutch-French cohort. *Ann Neurol,* 57**,** 505-12.

VAN DE WARRENBURG, B. P., SINKE, R. J., VERSCHUUREN-BEMELMANS, C. C., SCHEFFER, H., BRUNT, E. R., IPPEL, P. F., MAAT-KIEVIT, J. A., DOOIJES, D., NOTERMANS, N. C., LINDHOUT, D., KNOERS, N. V. & KREMER, H. P. 2002. Spinocerebellar ataxias in the Netherlands: prevalence and age at onset variance analysis. *Neurology,* 58**,** 702-8.

VAN DEN BROEK, W. J., NELEN, M. R., WANSINK, D. G., COERWINKEL, M. M., TE RIELE, H., GROENEN, P. J. & WIERINGA, B. 2002. Somatic expansion behaviour of the (CTG)n repeat in myotonic dystrophy knock-in mice is differentially affected by Msh3 and Msh6 mismatch-repair proteins. *Hum Mol Genet,* 11**,** 191-8.

VAN DER BURG, J. M., BJORKQVIST, M. & BRUNDIN, P. 2009. Beyond the brain: widespread pathology in Huntington's disease. *Lancet Neurol,* 8**,** 765-74.

VAN ENGELEN, B. & CONSORTIUM, O. 2015. Cognitive behaviour therapy plus aerobic exercise training to increase activity in patients with myotonic dystrophy type 1 (DM1) compared to usual care (OPTIMISTIC): study protocol for randomised controlled trial. *Trials,* 16**,** 224.

VAN VLIET, K. M., BLOUIN, V., BRUMENT, N., AGBANDJE-MCKENNA, M. & SNYDER, R. O. 2008. The role of the adeno-associated virus capsid in gene transfer. *Methods Mol Biol,* 437**,** 51-91.

VEITCH, N. J., ENNIS, M., MCABNEY, J. P., SHELBOURNE, P. F. & MONCKTON, D. G. 2007. Inherited CAG.CTG allele length is a major modifier of somatic mutation length variability in Huntington disease. *DNA Repair (Amst),* 6**,** 789-96.

VELAZQUEZ PEREZ, L., CRUZ, G. S., SANTOS FALCON, N., ENRIQUE ALMAGUER MEDEROS, L., ESCALONA BATALLAN, K., RODRIGUEZ LABRADA, R., PANEQUE HERRERA, M., LAFFITA MESA, J. M., RODRIGUEZ DIAZ, J. C., RODRIGUEZ, R. A., GONZALEZ ZALDIVAR, Y., COELLO ALMARALES, D., ALMAGUER GOTAY, D. & JORGE CEDENO, H. 2009. Molecular epidemiology of spinocerebellar ataxias in Cuba: insights into SCA2 founder effect in Holguin. *Neurosci Lett,* 454**,** 157-60.

VIDAL, R., CABALLERO, B., COUVE, A. & HETZ, C. 2011. Converging pathways in the occurrence of endoplasmic reticulum (ER) stress in Huntington's disease. *Curr Mol Med,* 11**,** 1-12.

VONSATTEL, J. P. 2008. Huntington disease models and human neuropathology: similarities and differences. *Acta Neuropathol,* 115**,** 55-69.

VONSATTEL, J. P., MYERS, R. H., STEVENS, T. J., FERRANTE, R. J., BIRD, E. D. & RICHARDSON, E. P., JR. 1985. Neuropathological classification of Huntington's disease. *J Neuropathol Exp Neurol,* 44**,** 559-77.

WANG, R., PERSKY, N. S., YOO, B., OUERFELLI, O., SMOGORZEWSKA, A., ELLEDGE, S. J. & PAVLETICH, N. P. 2014a. DNA repair. Mechanism of DNA interstrand cross-link processing by repair nuclease FAN1. *Science.* United States: American Association for the Advancement of Science.

WANG, R., PERSKY, N. S., YOO, B., OUERFELLI, O., SMOGORZEWSKA, A., ELLEDGE, S. J. & PAVLETICH, N. P. 2014b. Mechanism of DNA interstrand cross-link processing by repair nuclease FAN1. *Science* 346**,** 1127-1130.

WANG, S., LIU, K., XIAO, L., YANG, L., LI, H., ZHANG, F., LEI, L., LI, S., FENG, X., LI, A. & HE, J. 2016. Characterization of a novel DNA glycosylase from S. sahachiroi involved in the reduction and repair of azinomycin B induced DNA damage. *Nucleic Acids Res,* 44**,** 187-97.

WANG, W. 2007. Emergence of a DNA-damage response network consisting of Fanconi anaemia and BRCA proteins. *Nat Rev Genet,* 8**,** 735-48.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.

WARBY, S. C., GRAHAM, R. K. & HAYDEN, M. R. 2014. Huntington Disease.

WARDE-FARLEY, D., DONALDSON, S. L., COMES, O., ZUBERI, K., BADRAWI, R., CHAO, P., FRANZ, M., GROUIOS, C., KAZI, F., LOPES, C. T., MAITLAND, A., MOSTAFAVI, S., MONTOJO, J., SHAO, Q., WRIGHT, G., BADER, G. D. & MORRIS, Q. 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research,* 38**,** W214-W220.

WARDLE, M., MORRIS, H. R. & ROBERTSON, N. P. 2009. Clinical and genetic characteristics of non-Asian dentatorubral-pallidoluysian atrophy: A systematic review. *Mov Disord,* 24**,** 1636-40.

WARNER, J. P., BARRON, L. H. & BROCK, D. J. 1993. A new polymerase chain reaction (PCR) assay for the trinucleotide repeat that is unstable and expanded on Huntington's disease chromosomes. *Mol Cell Probes,* 7**,** 235-9.

WARNER, T. T., WILLIAMS, L. & HARDING, A. E. 1994. DRPLA in Europe. *Nat Genet,* 6**,** 225.

WATANABE, A., IKEJIMA, M., SUZUKI, N. & SHIMADA, T. 1996. Genomic organization and expression of the human MSH3 gene. *Genomics,* 31**,** 311-8.

WATANABE, H., TANAKA, F., DOYU, M., RIKU, S., YOSHIDA, M., HASHIZUME, Y. & SOBUE, G. 2000. Differential somatic CAG repeat instability in variable brain cell lineage in dentatorubral pallidoluysian atrophy (DRPLA): a laser-captured microdissection (LCM)-based analysis. *Hum Genet,* 107**,** 452-7.

WATASE, K., VENKEN, K. J., SUN, Y., ORR, H. T. & ZOGHBI, H. Y. 2003. Regional differences of somatic CAG repeat instability do not account for selective neuronal vulnerability in a knock-in mouse model of SCA1. *Hum Mol Genet,* 12**,** 2789-95.

WEISS, A., TRAGER, U., WILD, E. J., GRUENINGER, S., FARMER, R., LANDLES, C., SCAHILL, R. I., LAHIRI, N., HAIDER, S., MACDONALD, D., FROST, C., BATES, G. P., BILBE, G., KUHN, R., ANDRE, R. & TABRIZI, S. J. 2012. Mutant huntingtin fragmentation in immune cells tracks Huntington's disease progression. *J Clin Invest,* 122**,** 3731-6.

WEST, S. C. 2003. Molecular views of recombination proteins and their control. *Nat Rev Mol Cell Biol,* 4**,** 435-45.

WEXLER, N. S., LORIMER, J., PORTER, J., GOMEZ, F., MOSKOWITZ, C., SHACKELL, E., MARDER, K., PENCHASZADEH, G., ROBERTS, S. A., GAYAN, J., BROCKLEBANK, D., CHERNY, S. S., CARDON, L. R., GRAY, J., DLOUHY, S. R., WIKTORSKI, S., HODES, M. E., CONNEALLY, P. M., PENNEY, J. B., GUSELLA, J., CHA, J. H., IRIZARRY, M., ROSAS, D., HERSCH, S., HOLLINGSWORTH, Z., MACDONALD, M., YOUNG, A. B., ANDRESEN, J. M., HOUSMAN, D. E., DE YOUNG, M. M., BONILLA, E., STILLINGS, T., NEGRETTE, A., SNODGRASS, S. R., MARTINEZ-JAURRIETA, M. D., RAMOS-ARROYO, M. A., BICKHAM, J., RAMOS, J. S., MARSHALL, F., SHOULSON, I., REY, G. J., FEIGIN, A., ARNHEIM, N., ACEVEDO-CRUZ, A., ACOSTA, L., ALVIR, J., FISCHBECK, K., THOMPSON, L. M., YOUNG, A., DURE, L., O'BRIEN, C. J., PAULSEN, J., BRICKMAN, A., KRCH, D., PEERY, S., HOGARTH, P., HIGGINS, D. S., JR. & LANDWEHRMEYER, B. 2004a. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc Natl Acad Sci U S A,* 101**,** 3498-503.

WEXLER, N. S., LORIMER, J., PORTER, J., GOMEZ, F., MOSKOWITZ, C., SHACKELL, E., MARDER, K., PENCHASZADEH, G., ROBERTS, S. A., GAYAN, J., BROCKLEBANK, D., CHERNY, S. S., CARDON, L. R., GRAY, J., DLOUHY, S. R., WIKTORSKI, S., HODES, M. E., CONNEALLY, P. M., PENNEY, J. B., GUSELLA, J., CHA, J. H., IRIZARRY, M., ROSAS, D., HERSCH, S., HOLLINGSWORTH, Z., MACDONALD, M., YOUNG, A. B., ANDRESEN, J. M., HOUSMAN, D. E., DE YOUNG, M. M., BONILLA, E., STILLINGS, T., NEGRETTE, A., SNODGRASS, S. R., MARTINEZ-JAURRIETA, M. D., RAMOS-ARROYO, M. A., BICKHAM, J., RAMOS, J. S., MARSHALL, F., SHOULSON, I., REY, G. J., FEIGIN, A., ARNHEIM, N., ACEVEDO-CRUZ, A., ACOSTA, L., ALVIR, J., FISCHBECK, K., THOMPSON, L. M., YOUNG, A., DURE, L., O'BRIEN, C. J., PAULSEN, J., BRICKMAN, A., KRCH, D., PEERY, S., HOGARTH, P., HIGGINS, D. S., JR. & LANDWEHRMEYER, B. 2004b. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc Natl Acad Sci U S A.* United States.

WEYDT, P., PINEDA, V. V., TORRENCE, A. E., LIBBY, R. T., SATTERFIELD, T. F., LAZAROWSKI, E. R., GILBERT, M. L., MORTON, G. J., BAMMLER, T. K., STRAND, A. D., CUI, L., BEYER, R. P., EASLEY, C. N., SMITH, A. C., KRAINC, D., LUQUET, S., SWEET, I. R., SCHWARTZ, M. W. & LA SPADA, A. R. 2006. Thermoregulatory and metabolic defects in Huntington's disease transgenic mice implicate PGC-1alpha in Huntington's disease neurodegeneration. *Cell Metab,* 4**,** 349-62.

WHEELER, V. 1999. Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. *Human Molecular Genetics,* 8**,** 115-122.

WHEELER, V. C. 2003. Mismatch repair gene Msh2 modifies the timing of early disease in HdhQ111 striatum. *Human Molecular Genetics,* 12**,** 273-281.

WHEELER, V. C., AUERBACH, W., WHITE, J. K., SRINIDHI, J., AUERBACH, A., RYAN, A., DUYAO, M. P., VRBANAC, V., WEAVER, M., GUSELLA, J. F., JOYNER, A. L. & MACDONALD, M. E. 1999. Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. *Hum Mol Genet,* 8**,** 115-22.

WHEELER, V. C., LEBEL, L. A., VRBANAC, V., TEED, A., TE RIELE, H. & MACDONALD, M. E. 2003. Mismatch repair gene Msh2 modifies the timing of early disease in Hdh(Q111) striatum. *Hum Mol Genet,* 12**,** 273-81.

WHEELER, V. C., PERSICHETTI, F., MCNEIL, S. M., MYSORE, J. S., MYSORE, S. S., MACDONALD, M. E., MYERS, R. H., GUSELLA, J. F., WEXLER, N. S. & GROUP, U. S.-V. C. R. 2007. Factors associated with HD CAG repeat instability in Huntington disease. *J Med Genet,* 44**,** 695-701.

WHITNEY, A. R., DIEHN, M., POPPER, S. J., ALIZADEH, A. A., BOLDRICK, J. C., RELMAN, D. A. & BROWN, P. O. 2003. Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci U S A,* 100**,** 1896-901.

WIATR, K., SZLACHCIC, W. J., TRZECIAK, M., FIGLEROWICZ, M. & FIGIEL, M. 2018. Huntington Disease as a Neurodevelopmental Disorder and Early Signs of the Disease in Stem Cells. *Mol Neurobiol,* 55**,** 3351-3371.

WILD, E., MAGNUSSON, A., LAHIRI, N., KRUS, U., ORTH, M., TABRIZI, S. J. & BJÖRKQVIST, M. 2011. Abnormal peripheral chemokine profile in Huntington's disease. *PLoS Curr,* 3**,** RRN1231.

WILD, E. J., MUDANOHWO, E. E., SWEENEY, M. G., SCHNEIDER, S. A., BECK, J., BHATIA, K. P., ROSSOR, M. N., DAVIS, M. B. & TABRIZI, S. J. 2008. Huntington's disease phenocopies are clinically and genetically heterogeneous. *Mov Disord,* 23**,** 716-20.

WILD, E. J. & TABRIZI, S. J. 2007a. The differential diagnosis of chorea. *Pract Neurol,* 7**,** 360-73.

WILD, E. J. & TABRIZI, S. J. 2007b. Huntington's disease phenocopy syndromes. *Curr Opin Neurol,* 20**,** 681-7.

WILD, E. J. & TABRIZI, S. J. 2014. Targets for future clinical trials in Huntington's disease: What's in the pipeline? *Movement Disorders,* 29**,** 1434-1445.

WILLER, C. J., LI, Y. & ABECASIS, G. R. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics,* 26**,** 2190-1.

WILLIAMS, G. M. & SURTEES, J. A. 2015. MSH3 Promotes Dynamic Behavior of Trinucleotide Repeat Tracts In Vivo. *Genetics,* 200**,** 737-754.

WONG, L. J., ASHIZAWA, T., MONCKTON, D. G., CASKEY, C. T. & RICHARDS, C. S. 1995. Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. *Am J Hum Genet,* 56**,** 114-22.

WOOD, N. 2012. *Neurogenetics: A Guide for Clinicians*, Cambridge University Press.

WOOD-KACZMAR, A., GANDHI, S., YAO, Z., ABRAMOV, A. Y., MILJAN, E. A., KEEN, G., STANYER, L., HARGREAVES, I., KLUPSCH, K., DEAS, E., DOWNWARD, J., MANSFIELD, L., JAT, P., TAYLOR, J., HEALES, S., DUCHEN, M. R., LATCHMAN, D., TABRIZI, S. J. & WOOD, N. W. 2008. PINK1 is necessary for long term survival and mitochondrial function in human dopaminergic neurons. *PLoS One,* 3**,** e2455.

WRIGHT, G. E. B., COLLINS, J. A., KAY, C., MCDONALD, C., DOLZHENKO, E., XIA, Q., BEČANOVIĆ, K., SEMAKA, A., NGUYEN, C. M., TROST, B., RICHARDS, F., BIJLSMA, E. K., SQUITIERI, F., SCHERER, S. W., EBERLE, M. A., YUEN, R. K. C. & HAYDEN, M. R. 2019. Length of uninterrupted CAG repeats, independent of polyglutamine size, results in increased somatic instability and hastened age of onset in Huntington disease. 533414.

WYSS-CORAY, T. & ROGERS, J. 2012. Inflammation in Alzheimer disease-a brief review of the basic science and clinical literature. *Cold Spring Harb Perspect Med,* 2**,** a006346.

XIAO, X., MELTON, D. W. & GOURLEY, C. 2014. Mismatch repair deficiency in ovarian cancer -- molecular characteristics and clinical implications. *Gynecol Oncol,* 132**,** 506-12.

XU, H., ROSALES-REYNOSO, M. A., BARROS-NUNEZ, P. & PEPRAH, E. 2013. DNA repair/replication transcripts are down regulated in patients with Fragile X Syndrome. *BMC Res Notes,* 6**,** 90.

YAMAMOTO, H. & IMAI, K. 2015. Microsatellite instability: an update. *Arch Toxicol,* 89**,** 899-921.

YAN, P. X., HUO, Y. G. & JIANG, T. 2015. Crystal structure of human Fanconi-associated nuclease 1. *Protein Cell,* 6**,** 225-8.

YU, K., ROY, D., HUANG, F. T. & LIEBER, M. R. 2006. Detection and structural analysis of R-loops. *Methods Enzymol,* 409**,** 316-29.

ZATZ, M., PASSOS-BUENO, M. R., CERQUEIRA, A., MARIE, S. K., VAINZOF, M. & PAVANELLO, R. C. 1995. Analysis of the CTG repeat in skeletal muscle of young and adult myotonic dystrophy patients: when does the expansion occur? *Hum Mol Genet,* 4**,** 401-6.

ZENG, W., GILLIS, T., HAKKY, M., DJOUSSE, L., MYERS, R. H., MACDONALD, M. E. & GUSELLA, J. F. 2006. Genetic analysis of the GRIK2 modifier effect in Huntington's disease. *BMC Neurosci,* 7**,** 62.

ZHANG, B., GAITERI, C., BODEA, L. G., WANG, Z., MCELWEE, J., PODTELEZHNIKOV, A. A., ZHANG, C., XIE, T., TRAN, L., DOBRIN, R., FLUDER, E., CLURMAN, B., MELQUIST, S., NARAYANAN, M., SUVER, C., SHAH, H., MAHAJAN, M., GILLIS, T., MYSORE, J., MACDONALD, M. E., LAMB, J. R., BENNETT, D. A., MOLONY, C., STONE, D. J., GUDNASON, V., MYERS, A. J., SCHADT, E. E., NEUMANN, H., ZHU, J. & EMILSSON, V. 2013. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell,* 153**,** 707-20.

ZHANG, F., THORNHILL, S. I., HOWE, S. J., ULAGANATHAN, M., SCHAMBACH, A., SINCLAIR, J., KINNON, C., GASPAR, H. B., ANTONIOU, M. & THRASHER, A. J. 2007. Lentiviral vectors containing an enhancer-less ubiquitously acting chromatin opening element (UCOE) provide highly reproducible and stable transgene expression in hematopoietic cells. *Blood,* 110**,** 1448-57.

ZHANG, J. & WALTER, J. C. 2014. Mechanism and regulation of incisions during DNA interstrand cross-link repair. *DNA Repair (Amst),* 19**,** 135-42.

ZHANG, P., MO, J. Y., PEREZ, A., LEON, A., LIU, L., MAZLOUM, N., XU, H. & LEE, M. Y. 1999. Direct interaction of proliferating cell nuclear antigen with the p125 catalytic subunit of mammalian DNA polymerase delta. *J Biol Chem,* 274**,** 26647-53.

ZHAO, J., JAIN, A., IYER, R. R., MODRICH, P. L. & VASQUEZ, K. M. 2009. Mismatch repair and nucleotide excision repair proteins cooperate in the recognition of DNA interstrand crosslinks. *Nucleic Acids Res,* 37**,** 4420-9.

ZHAO, Q., XUE, X., LONGERICH, S., SUNG, P. & XIONG, Y. 2014. Structural insights into 5' flap DNA unwinding and incision by the human FAN1 dimer. *Nat Commun,* 5**,** 5726.

ZHAO, X.-N., KUMARI, D., GUPTA, S., WU, D., EVANITSKY, M., YANG, W. & USDIN, K. 2015a. Mutsβ generates both expansions and contractions in a mouse model of the Fragile X-associated disorders. *Hum. Mol. Genet.***,** ddv408.

ZHAO, X. N., KUMARI, D., GUPTA, S., WU, D., EVANITSKY, M., YANG, W. & USDIN, K. 2015b. Mutsbeta generates both expansions and contractions in a mouse model of the Fragile X-associated disorders. *Hum Mol Genet,* 24**,** 7087-96.

ZHAO, X. N., LOKANGA, R., ALLETTE, K., GAZY, I., WU, D. & USDIN, K. 2016. A MutSbeta-Dependent Contribution of MutSalpha to Repeat Expansions in Fragile X Premutation Mice? *PLoS Genet,* 12**,** e1006190.

ZHAO, X. N. & USDIN, K. 2018. FAN1 protects against repeat expansions in a Fragile X mouse model. *DNA Repair (Amst),* 69**,** 1-5.

ZHOU, W., OTTO, E. A., CLUCKEY, A., AIRIK, R., HURD, T. W., CHAKI, M., DIAZ, K., LACH, F. P., BENNETT, G. R., GEE, H. Y., GHOSH, A. K., NATARAJAN, S., THONGTHIP, S., VETURI, U., ALLEN, S. J., JANSSEN, S., RAMASWAMI, G., DIXON, J., BURKHALTER, F., SPOENDLIN, M., MOCH, H., MIHATSCH, M. J., VERINE, J., READE, R., SOLIMAN, H., GODIN, M., KISS, D., MONGA, G., MAZZUCCO, G., AMANN, K., ARTUNC, F., NEWLAND, R. C., WIECH, T., ZSCHIEDRICH, S., HUBER, T. B., FRIEDL, A., SLAATS, G. G., JOLES, J. A., GOLDSCHMEDING, R., WASHBURN, J., GILES, R. H., LEVY, S., SMOGORZEWSKA, A. & HILDEBRANDT, F. 2012. FAN1 mutations cause karyomegalic interstitial nephritis, linking chronic kidney failure to defective DNA damage repair. *Nat Genet,* 44**,** 910-5.

ZU, T., GIBBENS, B., DOTY, N. S., GOMES-PEREIRA, M., HUGUET, A., STONE, M. D., MARGOLIS, J., PETERSON, M., MARKOWSKI, T. W., INGRAM, M. A., NAN, Z., FORSTER, C., LOW, W. C., SCHOSER, B., SOMIA, N. V., CLARK, H. B., SCHMECHEL, S., BITTERMAN, P. B., GOURDON, G., SWANSON, M. S., MOSELEY, M. & RANUM, L. P. 2011. Non-ATG-initiated translation directed by microsatellite expansions. *Proc Natl Acad Sci U S A,* 108**,** 260-5.

ZUCCATO, C., BELYAEV, N., CONFORTI, P., OOI, L., TARTARI, M., PAPADIMOU, E., MACDONALD, M., FOSSALE, E., ZEITLIN, S., BUCKLEY, N. & CATTANEO, E. 2007. Widespread disruption of repressor element-1 silencing transcription factor/neuron-restrictive silencer factor occupancy at its target genes in Huntington's disease. *J Neurosci,* 27**,** 6972-83.

ZUCCATO, C., TARTARI, M., CROTTI, A., GOFFREDO, D., VALENZA, M., CONTI, L., CATAUDELLA, T., LEAVITT, B. R., HAYDEN, M. R., TIMMUSK, T., RIGAMONTI, D. & CATTANEO, E. 2003. Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat Genet,* 35**,** 76-83.

ZUHLKE, C., HELLENBROICH, Y., SCHAAFF, F., GEHLKEN, U., WESSEL, K., SCHUBERT, T., CERVOS-NAVARRO, J., PICKARTZ, H. & SCHWINGER, E. 1997. CAG repeat analyses in frozen and formalin-fixed tissues following primer extension preamplification for evaluation of mitotic instability of expanded SCA1 alleles. *Hum Genet,* 100**,** 339-44.

ZWILLING, D., HUANG, S. Y., SATHYASAIKUMAR, K. V., NOTARANGELO, F. M., GUIDETTI, P., WU, H. Q., LEE, J., TRUONG, J., ANDREWS-ZWILLING, Y., HSIEH, E. W., LOUIE, J. Y., WU, T., SCEARCE-

LEVIE, K., PATRICK, C., ADAME, A., GIORGINI, F., MOUSSAOUI, S., LAUE, G., RASSOULPOUR, A., FLIK, G., HUANG, Y., MUCHOWSKI, J. M., MASLIAH, E., SCHWARCZ, R. & MUCHOWSKI, P. J. 2011. Kynurenine 3-monooxygenase inhibition in blood ameliorates neurodegeneration. *Cell,* 145**,** 863-74.